cemmap
centre for microdata methods and practice

Nonseparable sample selection models with censored selection rules

Ivan Fernández-Val
Aico van Vuuren
Francis Vella

The Institute for Fiscal Studies
Department of Economics, UCL

cemmap working paper CWP10/18

E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL
An ESRC Research Centre

Nonseparable Sample Selection Models with Censored Selection Rules *

Ivan Fernández-Val† Aico van Vuuren‡ Francis Vella§

January 25, 2018

Abstract

We consider identification and estimation of nonseparable sample selection models with censored selection rules. We employ a control function approach and discuss different objects of interest based on (1) local effects conditional on the control function, and (2) global effects obtained from integration over ranges of values of the control function. We provide conditions under which these objects are appropriate for the total population. We also present results regarding the estimation of counterfactual distributions. We derive conditions for identification for these different objects and suggest strategies for estimation. We also provide the associated asymptotic theory. These strategies are illustrated in an empirical investigation of the determinants of female wages and wage growth in the United Kingdom.

Keywords: Sample selection, nonseparable models, control function, quantile and distribution regression

JEL-codes: C14, C21, C24

*We thank Costas Meghir for providing the data, and Stéphane Bonhomme, Ivana Komunjer and Sami Stouli for useful comments.

†Boston University.

‡University of Gothenburg.

§Georgetown University.

1 Introduction

This paper considers a nonseparable sample selection model with a censored selection rule. The most common example is a selection rule with censoring at zero, also referred to in the parametric setting as tobit type 3, although other forms of censored selection rules are permissible. A leading empirical example is estimating the determinants of wages when workers report working hours rather than the binary work/not work decision. An important feature of the model, beyond the relaxation of distributional assumptions, is the inherent heterogeneity facilitated through nonseparability. Our approach is to account for selection via an appropriately constructed control function. We propose a three step estimation procedure which first employs the distribution regression of Foresi and Peracchi (1995) and Chernozhukov et al. (2013) to compute the appropriate control function. The second step is estimated either by least squares, distribution or quantile regression employing the estimated control function. The primary estimands of interest are obtained in the third step as functionals of the second step and control function estimates.

Our paper contributes to the growing literatures on nonseparable models with endogeneity (see, for example, Chesher 2003, Ma and Koenker 2006, Florens et al. 2008, Imbens and Newey 2009, Jun 2009 and Masten and Torgovitsky, 2014) and nonseparable sample selection models (for example, Newey 2007). An important contribution is our focus on the identification and estimation of local effects. While Newey (2007) considered the distribution of the outcome variable conditional on selection, we provide statements regarding the outcome variable distribution conditional on specific values of the control function. This local approach to identification is popular in many contexts (see, for example, Chesher 2003, and Heckman and Vytlacil 2005). We show that for any population observation that has a positive probability of being selected, selection is irrelevant for the distribution of the outcome variable conditional on the control function. Hence, we can estimate certain objects of interest that are appropriate for the whole population conditional on the value of the control function. We can also estimate global objects by integrating over the distribution of the control function in the selected or entire population. However, we highlight that these global objects require strong support assumptions on the explanatory variables which may be difficult to satisfy in empirical applications.

Accordingly, we also consider global effects "on the treated" that are identified under weaker assumptions. In addition to defining and providing estimators of these global and local effects, we provide their associated asymptotic theory.

This paper is also related to the literature on quantile selection models. Arellano and Bonhomme (2017) addressed selection by modeling the copula of the error terms in the outcome and selection equations. The most important distinction to this paper is that they consider the conventional binary, rather than a censored, selection equation. Thus we require more information about the selection process. However this has the advantage that one can consider local effects conditional on the control function which are identified under weaker conditions.

The following section outlines the model and some related literature. Section 3 defines the control function and provides identification results regarding the objects of interest in the model. Section 4 provides estimators of these objects and discusses inference. Section 5 illustrates some of our estimands focusing on the determinants of wages and wage growth for working women in the United Kingdom.

2 Model

The model has the following structure:

$$Y = g(X, \varepsilon) \text{ if } C > 0, \tag{2.1}$$

$$C = \max\left(h(Z, \eta), 0\right), \tag{2.2}$$

where Y and C are observable random variables, and X and Z are vectors of observable explanatory variables. The set of variables included in X is a subset of the set of variables included in Z. In principle, we do not need to impose an exclusion restriction on Z with respect to the elements of X, although our identification assumptions will be more plausible under such a restriction. The functions g and h are unknown and ε and η are respectively a vector and a scalar of potentially mutually dependent unobservables. We shall impose restrictions on the stochastic properties of these unobservables. The primary objective is to estimate functionals related to g noting that Y is only observed when C is above some known threshold normalized

to be zero. The non observability of Y for specific values of C induces the possibility of selection bias. We shall refer to (2.1) as the outcome equation and (2.2) as the selection equation.

The model is a nonparametric and nonseparable representation of the tobit type-3 model and is a variant of the Heckman (1979) selection model. It was initially examined in a fully parametric setting, imposing additivity and normality, and estimated by maximum likelihood (see Amemiya, 1978, 1979). Vella (1993) provided a two-step estimator based on estimating the generalized residual from the selection equation and including it as a control function in the outcome equation. Honoré et al. (1997), Chen (1997) and Lee and Vella (2006) relaxed the model's distributional assumptions, but imposed an index restriction and separability of the error terms in each equation.

The model can be extended in several directions. For example, the selection variable C could be censored in a number of ways provided there are some region(s) for which it is continuously observed. This allows for top, middle and/or bottom censoring. Also, although we do not consider it explicitly here, our approach is applicable when the outcome variable Y is also censored. For example:

$$Y = \max(g(X, \varepsilon), 0) \text{ if } C > 0.$$

The model can also be extended to include C in the outcome equation as explanatory variable provided that there is an exclusion restriction in Z with respect to X. This extension, which corresponds to the triangular system of Imbens and Newey (2009) with censoring in the first stage equation, is not considered here as it is not relevant for our empirical application.

We highlighted above that we follow a local approach to identification such as proposed by Heckman and Vytlacil (2005) who consider a binary treatment/selection rule and a separable selection equation. While our focus is also, in part, on local effects our model differs with respect to the selection rule and the possible presence of nonseparability.

3 Identification of objects of interest

We account for selection bias through the use of an appropriately constructed control function. Accordingly, we first establish the existence of such a function for this model and then define some objects of interest incorporated in (2.1)-(2.2).

Let $\perp\!\!\!\perp$ denote stochastic independence. We begin with the following assumption:

Assumption 1 (Control Function) $(\varepsilon, \eta) \perp\!\!\!\perp Z$, *$\eta$ is a continuously distributed random variable with strictly increasing CDF on the support of η, and $t \mapsto h(Z, t)$ is strictly increasing a.s.*

This assumption allows for endogeneity between X and ε in the selected population with $C > 0$, since in general ε and η are dependent, *i.e.* $\varepsilon \not\perp\!\!\!\perp X \mid C > 0$. The monotonicity assumption allows a non-monotonic relationship between ε and C because ε and η are allowed to be non-monotonically dependent. Under Assumption 1, we can normalize the distribution of η to be uniform on $[0, 1]$ without loss of generality (Matzkin, 2003).[1]

The following result shows the existence of a control function for the selected population in this setting. That is, there is a function of the observable data such that once it is conditioned upon, the unobservable component is independent of the explanatory variables in the outcome equation for the selected population. Let $V := F_C(C \mid Z)$ where $F_C(\cdot \mid z)$ denotes the CDF of C conditional on $Z = z$.

Lemma 1 (Existence of Control Function) *Under the model in (2.1)-(2.2) and Assumption 1:*

$$\varepsilon \perp\!\!\!\perp Z \mid V, C > 0.$$

All proofs are provided in the Appendix. The intuition behind Lemma 1 is based on three observations. First, $V = \eta$ when $C > 0$, so that conditioning on V is identical to conditioning on η in the selected population. Second, conditioning on Z and η makes selection, *i.e.* $C > 0$, deterministic. Therefore, the distribution of ε, conditional on Z and η, does not depend on the condition that $C > 0$. The

[1] Indeed if $t \mapsto h(z, t)$ is strictly increasing, and η is continuously distributed with $\eta \sim F_\eta$, then $\tilde{h}(z, \tilde{\eta}) = h(z, F_\eta(\tilde{\eta}))$ is such that $t \mapsto \tilde{h}(z, t)$ is strictly increasing and $\tilde{\eta} \sim U(0, 1)$.

final observation, namely our assumption that $(\varepsilon, \eta) \perp\!\!\!\perp Z$, is sufficient to prove the Lemma.

We consider two classes of objects which are interesting for econometric inference. These are: (1) local effects conditional on the value of the control function, and (2) global effects based on integration over the control function.

3.1 Local effects

We consider local effects on Y for given values of X conditional on the control function V. Let $\mathcal{Z}$, $\mathcal{X}$, and $\mathcal{V}$ denote the marginal supports of Z, X, and V in the selected population, respectively. We start by introducing the set $\mathcal{XV}$, the joint support of X and V in the selected population.

Definition 1 (Identification set) *Define*

$$\mathcal{XV} := \{(x, v) \in \mathcal{X} \times \mathcal{V} : h(z, v) > 0, z \in \mathcal{Z}(x)\},$$

where $\mathcal{Z}(x) = \{z \in \mathcal{Z} : x \subseteq z\}$, i.e. *the set of values of* Z *with the component* $X = x$.

Depending on the values of (X, η), we can classify the units of observation into 3 groups: (1) always selected units when $h(z, t) > 0$ for all $z \in \mathcal{Z}(x)$, (2) switchers when $h(z, t) > 0$ for some $z \in \mathcal{Z}(x)$ and $h(z, t) \leq 0$ for some $z \in \mathcal{Z}(x)$, and (3) never selected units when $h(z, t) \leq 0$ for all $z \in \mathcal{Z}(x)$. The set $\mathcal{XV}$ only includes always selected units and switchers, *i.e.* units with (X, V) such that they are observed for some values of Z. When $X = Z$ there are no switchers because the set $\mathcal{Z}(x)$ is a singleton. Otherwise the size of the set $\mathcal{XV}$ increases with the support of the excluded variables and their strength in the selection equation.

We now define the first local effect, the local average structural function.

Definition 2 (LASF) *The local average structural function (LASF) at* (x, v) *is:*

$$\mu(x, v) = \mathbb{E}(g(x, \varepsilon) \mid V = v).$$

The LASF gives the expected value of the potential outcome $g(x, \varepsilon)$ obtained by fixing X at x conditional on $V = v$ for the entire population. It is useful for

measuring the effect of X on the mean of Y. For example, the average treatment effect of changing X from x_0 to x_1 conditional on $V = v$ is

$$\mu(x_1, v) - \mu(x_0, v).$$

The following result shows that $\mu(x, v)$ is identified for all $(x, v) \in \mathcal{XV}$.

Theorem 1 (Identification of LASF) *Under the model (2.1)-(2.2), Assumption 1 and $\mathbb{E}|Y| < \infty$, for $(x, v) \in \mathcal{XV}$,*

$$\mu(x, v) = \mathbb{E}(Y \mid X = x, V = v, C > 0). \tag{3.1}$$

According to Theorem 1, the LASF is identical to the expected value of the outcome variable conditional on $(X, V) = (x, v)$ in the selected population. The proof of this theorem is based on Assumption 1 that allows for the LASF to be conditional on the outcome of $(Z, V) = (z, v)$. Since $(x, v) \in \mathcal{XV}$, there is a $z \in \mathcal{Z}(x)$ such that $h(z, v) > 0$ and hence the expected mean outcome of $g(x, \varepsilon)$ conditional on $V = v$ for the total sample, *i.e.* the LASF, is the same as that mean outcome for the selected sample. That is, selection is irrelevant for the distribution of the outcome variable conditional on the control function. This mean outcome is equal to the conditional expectation in the selected population, which is a function of the data distribution and is hence identified.

When X is continuous and $x \mapsto g(x, \varepsilon)$ is differentiable a.s., we can consider the average derivative of $g(x, \varepsilon)$ with respect to x conditional on the control function.

Definition 3 (LADF) *The local average derivative function (LADF) at (x, v) is:*

$$\delta(x, v) = \mathbb{E}[\partial_x g(x, \varepsilon) \mid V = v], \quad \partial_x := \partial/\partial x. \tag{3.2}$$

The LADF is the first-order derivative of the LASF with respect to x, provided that we can interchange differentiation and integration in (3.2). This is made formal in the next corollary which shows that the LADF is identified for all $(x, v) \in \mathcal{XV}$.

Corollary 1 (Identification of LADF) *Assume that for all $x \in \mathcal{X}$, $g(x, \varepsilon)$ is continuously differentiable in x a.s., $E[|g(x, \varepsilon)|] < \infty$, and $E[|\partial_x g(x, \varepsilon)|] < \infty$.*

Under the conditions of Theorem 1, for $(x, v) \in \mathcal{XV}$,

$$\delta(x, v) = \partial_x \mu(x, v) = \partial_x \mathbb{E}(Y \mid X = x, V = v, C > 0).$$

The local effects extend in a straightforward manner to distributions and quantiles.

Definition 4 (LDSF and LQSF) *The local distribution structural function (LDSF) at* (y, x, v) *is:*

$$G(y, x, v) = \mathbb{E}[1\{g(x, \varepsilon) \leq y\} \mid V = v].$$

The local quantile structural function (LQSF) at (τ, x, v) *is:*

$$q(\tau, x, v) := \inf\{y \in \mathbb{R} : G(y, x, v) \geq \tau\}.$$

The LDSF is the distribution function of the potential outcome $g(x, \varepsilon)$ conditional on the value of the control function for the entire population. The LQSF is the left-inverse function of $y \mapsto G(y, x, v)$ and corresponds to quantiles of $g(x, \varepsilon)$. Differences of the LQSF across levels of x correspond to quantile effects conditional on V for the entire population. For example, the τ-quantile treatment effect of changing X from x_0 to x_1 is

$$q(\tau, x_1, v) - q(\tau, x_0, v).$$

The identification of the LDSF follows by the same argument as the identification of the LASF, replacing $g(x, \varepsilon)$ (as in Definition 2) by $1\{g(x, \varepsilon) \leq y\}$ and Y (as in equation (3.1)) by $1\{Y \leq y\}$. Thus, under Assumption 1, for $(x, v) \in \mathcal{XV}$,

$$\mathbb{E}[1\{g(x, \varepsilon) \leq y\} \mid V = v] = F_{Y|X,V,C>0}(y \mid x, v).$$

The LQSF is then identified by the left-inverse function of $y \mapsto F_{Y|X,V,C>0}(y \mid x, v)$, the conditional quantile function $\tau \mapsto \mathbb{Q}_Y[\tau \mid X = x, V = v, C > 0]$, *i.e.*, for $(x, v) \in \mathcal{XV}$,

$$q(\tau, x, v) = \mathbb{Q}_Y[\tau \mid X = x, V = v, C > 0].$$

We also consider the derivative of $q(\tau, x, v)$ with respect to x and call it the local quantile derivative function (LQDF). This object corresponds to the average derivative of $g(x, \varepsilon)$ with respect to x at the quantile $q(\tau, x, v)$ conditional on $V = v$

under suitable regularity conditions; see Hoderlein and Mammen (2011). Thus, for $(\tau, x, v) \in [0,1] \times \mathcal{XV}$,

$$\delta_\tau(x, v) := \partial_x q(\tau, x, v) = \mathbb{E}[\partial_x g(x, \varepsilon) \mid V = v, g(x, \varepsilon) = q(\tau, x, v)].$$

By an analogous argument to Corollary 1, the LQDF is identified at $(\tau, x, v) \in [0,1] \times \mathcal{XV}$ by:

$$\delta_\tau(x, v) = \partial_x \mathbb{Q}_Y[\tau \mid X = x, V = v, C > 0],$$

provided that $x \mapsto \mathbb{Q}_Y[\tau \mid X = x, V = v, C > 0]$ is differentiable and other regularity conditions hold.

Remark 1 (Exclusion restrictions) *The identification of local effects does not explicitly require exclusion restrictions in Z with respect to X although the size of the identification set $\mathcal{XV}$ depends on such restrictions. For example, if $h(z, \eta) = z + \Phi^{-1}(\eta)$ where Φ is the standard normal distribution and $X = Z$, then $\mathcal{XV} = \{(x, v) \in \mathcal{X} \times \mathcal{V}] : x \leq -\Phi^{-1}(v)\} \subset \mathcal{X} \times \mathcal{V}$; whereas if $h(z, \eta) = x + z_1 + \Phi^{-1}(\eta)$ for $Z = (X, Z_1)$, then $\mathcal{XV} = \{(x, v) \in \mathcal{X} \times \mathcal{V} : x \leq -\Phi^{-1}(v) - z_1, z_1 \in \mathcal{Z}(x)\}$, such that $\mathcal{XV} = \mathcal{X} \times \mathcal{V}$ if Z_1 is independent of X and supported in $\mathbb{R}$.*

3.2 Global effects

We expand our set of estimands by examining the global counterparts of the local effects obtained by integration over the control function in the selected population. A typical global effect at $x \in \mathcal{X}$ is:

$$\theta_S(x) = \int \theta(x, v) dF_{V|C>0}(v), \tag{3.3}$$

where $\theta(x, v)$ can be any of the local objects defined above and $F_{V|C>0}(v)$ is the distribution of V in the selected population. Identification of $\theta_S(x)$ requires identification of $\theta(x, v)$ over $\mathcal{V}$, the support of V in the selected population.

For example, the average structural function (ASF),

$$\mu_S(x) := \mathbb{E}[g(x, \varepsilon) \mid C > 0],$$

gives the average of the potential outcome $g(x, \varepsilon)$ in the selected population. By the law of iterated expectations, this is a special case of the global effect (3.3) with $\theta(x, v) = \mu(x, v)$, the LASF. The average treatment effect of changing X from x_0 to x_1 in the selected population is

$$\mu_S(x_1) - \mu_S(x_0).$$

Similarly, one can consider the distribution structural function (DSF) in the selected population as in Newey (2007) , *i.e:*

$$G_S(y, x) := \mathbb{E}[1\{g(x, \varepsilon) \leq y\} \mid C > 0],$$

which gives the distribution of the potential outcome $g(x, \varepsilon)$ at y in the selected population. This is also a special case of the global effect (3.3) with $\theta(x, v) = G(y, x, v)$. We can then construct the quantile structural function (QSF) in the selected population as the left-inverse of $y \mapsto G_S(y, x)$, that is:

$$q_S(\tau, x) := \inf\{y \in \mathbb{R} : G_S(y, x) \geq \tau\}.$$

The QSF gives the quantiles of $g(x, \varepsilon)$. Unlike $G_S(y, x)$, $q_S(\tau, x)$ cannot be obtained by integration of the corresponding local effect, $q(\tau, x, v)$, because we cannot interchange quantiles and expectations. The τ-quantile treatment effect of changing X from x_0 to x_1 in the selected population is

$$q_S(\tau, x_1) - q_S(\tau, x_0).$$

Global counterparts of the LADF and LQSF are obtained by taking derivatives of $\mu_S(x)$ and $q_S(\tau, x)$ with respect to x.

As in Newey (2007), identification of the global effects in the selected population requires a condition on the support of the control function. Let $\mathcal{V}(x)$ denote the support of V conditional on $X = x$, *i.e.* $\mathcal{V}(x) := \{v \in \mathcal{V} : (x, v) \in \mathcal{XV}\}$.

Assumption 2 (Common Support) $\mathcal{V}(x) = \mathcal{V}$.

The main implication of common support is the identification of $\theta(x)$ from the

identification of $\theta(x, v)$ in $v \in \mathcal{V}(x) = \mathcal{V}$. Assumption 2 is only plausible under exclusion restrictions on Z with respect to X; see the example in Remark 1.

We now establish the identification of the typical global effect (3.3).

Theorem 2 (Identification of Global Effects) *If $\theta(x, v)$ is identified for all $(x, v) \in \mathcal{X}\mathcal{V}$, then $\theta_S(x)$ is identified for all $x \in \mathcal{X}$ that satisfy Assumption 2.*

We can now apply this result to show identification of global effects in the selected population, because under Assumption 1 the local effects are identified over $\mathcal{X}\mathcal{V}$, which is the support of (X, V) in the selected population.

Remark 2 (Global Effects in the Entire Population) *The effects in the selected population generally differ from the effects in the entire population, except under the additional support condition:*

$$\mathcal{V} = (0, 1), \tag{3.4}$$

which imposes that the control function is fully supported in the selected population. This condition requires an excluded variable in Z with sufficient variation to make $h(Z, \eta) > 0$ for any $\eta \in [0, 1]$ by an identification at infinity argument.

3.3 Global effects on the treated and average derivatives

Assumption 2 might be too restrictive for empirical applications where an excluded variable with large support is not available. Without this assumption the global effects are not point identified, but can be bounded following a similar approach to Imbens and Newey (2009). We consider instead the alternative generic global effect:

$$\theta_S(x \mid x_0) = \int \theta(x, v) dF_{V|X,C>0}(v \mid x_0), \tag{3.5}$$

which is point identified under weaker support conditions than (3.3). Examples of (3.5) include the ASF conditional on $X = x_0$ in the selected population,

$$\mu_S(x \mid x_0) = \mathbb{E}[g(x, \varepsilon) \mid X = x_0, C > 0],$$

which is a special case of (3.5) with $\theta(x, v) = \mu(x, v)$. This ASF measures the mean of the potential outcome $g(x, \varepsilon)$ for the selected individuals with $X = x_0$, and is useful to construct the average treatment effect on the treated of changing X from x_0 to x_1

$$\mu_S(x_1 \mid x_0) - \mu_S(x_0 \mid x_0).$$

The object in (3.5) is identified in the selected population under the following support condition:

Assumption 3 (Weak Common Support) $\mathcal{V}(x) \supseteq \mathcal{V}(x_0)$.

Assumption 3 is weaker than Assumption 2 because $\mathcal{V}(x_0) \subseteq \mathcal{V}$. In particular, if the selection equation (2.2) is monotone in X and $\mathcal{X}$ is bounded from below, Assumption 3 is satisfied by setting x_0 lower than x.

We define the τ-quantile treatment on the treated as:

$$q_S(\tau, x_1 \mid x_0) - q_S(\tau, x_0 \mid x_0),$$

where $q_S(\tau, x \mid x_0)$ is the left-inverse of the DSF conditional on $X = x_0$ in the selected population,

$$G_S(y, x \mid x_0) := \mathbb{E}[1\{g(x, \varepsilon) \leq y\} \mid X = x_0, C > 0],$$

which is a special case of the effect (3.5) with $\theta(x, v) = G(y, x, v)$.

We now establish the identification of the typical global effect (3.5).

Theorem 3 (Identification of Global effects on the Treated) *If $\theta(x, v)$ is identified for all $(x, v) \in \mathcal{X}\mathcal{V}$, then $\theta_S(x \mid x_0)$ is identified for all $x \in \mathcal{X}$ that satisfy Assumption 3.*

We can define global objects in the selected population that are identified without a common support assumption when X is continuous and $x \mapsto g(x, \cdot)$ is differentiable. One example is the average derivative conditional on $X = x$ in the selected population:

$$\delta_S(x) = \mathbb{E}[\delta(x, V) \mid X = x, C > 0],$$

which is a special case of the effect (3.5) with $\theta(x, v) = \delta(x, v)$ and $x_0 = x$. This object is point identified in the selected population under Assumption 1 because the integral is over $\mathcal{V}(x)$, the support of V conditional on $X = x$ in the selected population. Another example is the average derivative in the selected population:

$$\delta_S = \mathbb{E}[\delta(X, V) \mid C > 0],$$

which is point identified under Assumption 1 because the integral is over $\mathcal{XV}$, the support of (X, V) in the selected population. This is a special case of the generic global effect

$$\theta_S = \int \theta(x, v) dF_{XV|C>0}(x, v). \tag{3.6}$$

3.4 Counterfactual distributions

We also consider linear functionals of the global effects including counterfactual distributions constructed by integration of the DSF with respect to different distributions of the explanatory variables and control function. These counterfactual distributions are useful for performing wage decompositions and other counterfactual analyses (e.g., DiNardo et al., 1996, Chernozhukov et al., 2013, Firpo et al., 2011, and Arellano and Bonhomme, 2017).

We focus on functionals in the selected population. To simplify the notation, we use a superscript s to denote these functionals, instead of explicitly conditioning on $C > 0$. The basis of the decompositions is the following expression for the observed distribution of Y:

$$G_Y^s(y) = \int F_{Y|Z,V}^s(y \mid z, v) dF_{Z,V}^s(z, v). \tag{3.7}$$

We show in the Appendix that (3.7) can be rewritten as:

$$G_Y^s(y) = \frac{\int G(y, x, v) 1(h(z, v) > 0) dF_{Z,V}(z, v)}{\int 1(h(z, v) > 0) dF_{Z,V}(z, v)}. \tag{3.8}$$

We construct counterfactual distributions by combining the component distributions G and $F_{Z,V}$ as well as the selection rule h from different populations that can correspond to different time periods or demographic groups. Thus, let G_t and F_{Z_k,V_k} denote the distributions in groups t and k, and h_r denote the selection rule in group

r. Then, the counterfactual distribution of Y when G is as in group t, $F_{Z,V}$ is as in group k, while the selection rule is identical to group r, is:

$$G^s_{Y_{\langle t|k,r\rangle}}(y) := \frac{\int G_t(y,x,v)1(h_r(z,v)>0)dF_{Z_k,V_k}(z,v)}{\int 1(h_r(z,v)>0)dF_{Z_k,V_k}(z,v)}. \tag{3.9}$$

Note that under this definition the observed distribution in group t is $G^s_{Y_{\langle t|t,t\rangle}}$. A sufficient condition for nonparametric identification is that $\mathcal{ZV}_k \subseteq \mathcal{ZV}_r \subseteq \mathcal{ZV}_t$, which guarantees that G_t and h_r are identified for all combinations of z and v over which we integrate. By monotonicity of $v \mapsto h(z,v)$, the condition $h_r(z,v) > 0$ is equivalent to

$$v > F_{C_r|Z}(0 \mid z), \tag{3.10}$$

where $F_{C_r|Z}$ is the distribution of C conditional on Z in group r. Note that the identification condition $\mathcal{ZV}_k \subseteq \mathcal{ZV}_r$ can be weakened to $\mathcal{Z}_k \subseteq \mathcal{Z}_r$, and $\mathcal{ZV}_r \subseteq \mathcal{ZV}_t$ to $\mathcal{XV}_r \subseteq \mathcal{XV}_t$, which is more plausible in the presence of exclusion restrictions in X with respect to Z.

We can decompose the difference in the observed distribution between group 1 and 0 using counterfactual distributions:

$$G^s_{Y_{\langle 1|1,1\rangle}} - G^s_{Y_{\langle 0|0,0\rangle}} = \underbrace{[G^s_{Y_{\langle 1|1,1\rangle}} - G^s_{Y_{\langle 1|1,0\rangle}}]}_{(1)} + \underbrace{[G^s_{Y_{\langle 1|1,0\rangle}} - G^s_{Y_{\langle 1|0,0\rangle}}]}_{(2)} + \underbrace{[G^s_{Y_{\langle 1|0,0\rangle}} - G^s_{Y_{\langle 0|0,0\rangle}}]}_{(3)}, \tag{3.11}$$

where (1) is a selection effect due to the change in the selection rule given the distribution of the explanatory variables and the control function, (2) is a composition effect due to the change in the distribution of the explanatory variables and the control function, and (3) is a structure effect due to the change in the conditional distribution of the outcome given the explanatory variables and control function.

4 Estimation and inference

The effects of interest are all identified by functionals of the distribution of the observed variables and the control function in the selected population. The control function is the distribution of the censoring variable C conditional on all the explana-

tory variables Z. We propose a multistep semiparametric method based on least squares, distribution and quantile regressions to estimate the effects. The reduced form specifications used in each step can be motivated by parametric restrictions on the model (2.1)–(2.2). We refer to Chernozhukov et al. (2017) for examples of such restrictions.

Throughout this section, we assume that we have a random sample of size n, $\{(Y_i * 1(C_i > 0), C_i, Z_i)\}_{i=1}^n$, of the random variables $(Y * 1(C > 0), C, Z)$, where $Y * 1(C > 0)$ indicates that Y is observed only when $C > 0$.

4.1 Step 1: Estimation of the control function

We estimate the control function using logistic distribution regression (Foresi and Peracchi, 1995, and Chernozhukov et al., 2013). More precisely, for every observation in the selected sample, we set:

$$\widehat{V}_i = \Lambda(R_i^{\mathrm{T}} \widehat{\pi}(C_i)), \quad R_i := r(Z_i), \quad i = 1, \ldots, n, C_i > 0,$$

where, for $c \in \mathcal{C}_n$, the empirical support of C,

$$\widehat{\pi}(c) = \arg\max_{\pi \in \mathbb{R}^{d_r}} \sum_{i=1}^{n} \left[1\{C_i \leq c\} \log \Lambda(R_i^{\mathrm{T}} \pi)) + 1\{C_i > c\} \log \Lambda(-R_i^{\mathrm{T}} \pi) \right],$$

Λ is the logistic distribution, and $r(z)$ is a d_r-dimensional vector of transformations of z with good approximating properties such as polynomials, B-splines and interactions.

4.2 Step 2: Estimation of local objects

We can estimate the local average, distribution and quantile structural functions using flexibly parametrized least squares, distribution and quantile regressions, where we replace the control function by its estimator from the previous step.

For reasons explained in Section 4.6, our estimation method is based on a trimmed sample with respect to the censoring variable C. Therefore, we introduce

the following trimming indicator among the selected sample

$$T = 1(C \in \overline{\mathcal{C}})$$

where $\overline{\mathcal{C}} = (0, \overline{c}]$ for some $0 < \overline{c} < \infty$, such that $P(T = 1) > 0$.

The estimator of the LASF is $\widehat{\mu}(x, v) = w(x, v)^{\mathrm{T}}\widehat{\beta}$, where $w(x, v)$ is a d_w-dimensional vector of transformations of (x, v) with good approximating properties, and $\widehat{\beta}$ is the ordinary least squares estimator:[2]

$$\widehat{\beta} = \left[\sum_{i=1}^{n} \widehat{W}_i \widehat{W}_i^{\mathrm{T}} T_i\right]^{-1} \sum_{i=1}^{n} \widehat{W}_i Y_i T_i, \quad \widehat{W}_i := w(X_i, \widehat{V}_i).$$

The estimator of the LDSF is $\widehat{G}(y, x, v) = \Lambda(w(x, v)^{\mathrm{T}}\widehat{\beta}(y))$, where $\widehat{\beta}(y)$ is the logistic distribution regression estimator:

$$\widehat{\beta}(y) = \arg\max_{b \in \mathbb{R}^{d_w}} \sum_{i=1}^{n} \left[1\{Y_i \leq y\} \log \Lambda(\widehat{W}_i^{\mathrm{T}} b)) + 1\{Y_i > y\} \log \Lambda(-\widehat{W}_i^{\mathrm{T}} b))\right] T_i.$$

Similarly, the estimator of the LQSF is $\widehat{q}(\tau, x, v) = w(x, v)^{\mathrm{T}}\widehat{\beta}(\tau)$, where $\widehat{\beta}(\tau)$ is the Koenker and Bassett (1978) quantile regression estimator

$$\widehat{\beta}(\tau) = \arg\min_{b \in \mathbb{R}^{d_w}} \sum_{i=1}^{n} \rho_\tau(Y_i - \widehat{W}_i^{\mathrm{T}} b) T_i.$$

Estimators of the local derivatives are obtained by taking derivatives of the estimators of the local structural functions. Thus, the estimator of the LADF is:

$$\widehat{\delta}(x, v) = \partial_x w(x, v)^{\mathrm{T}}\widehat{\beta},$$

and the estimator of the LQDF is:

$$\widehat{\delta}_\tau(x, v) = \partial_x w(x, v)^{\mathrm{T}}\widehat{\beta}(\tau).$$

[2] An alternative approach is to follow Jun (2009) and Masten and Torgovitsky (2014). These papers acknowledge that with an index restriction the parameters of interest can be estimated in the presence of a control function by estimation over subsamples for which the control function has a *similar* value. While each of these papers considers a random coefficients model with endogeneity their approach is applicable here.

4.3 Step 3: Estimation of global effects

We obtain estimators of the generic global effects by approximating the integrals over the control function by averages of the estimated local effects evaluated at the estimated control function. The estimator of the effect (3.3) is

$$\widehat{\theta}_S(x) = \sum_{i=1}^{n} T_i \widehat{\theta}(x, \widehat{V}_i) / \sum_{i=1}^{n} T_i.$$

This yields the estimators of the ASF for $\widehat{\theta}(x, v) = \widehat{\mu}(x, v)$ and DSF at y for $\widehat{\theta}(x, v) = \widehat{G}(y, x, v)$. The estimator of the QSF is then obtained by inversion of the estimator of the DSF.[3] We form an estimator of the effect (3.5) as

$$\widehat{\theta}_S(x \mid x_0) = \sum_{i=1}^{n} T_i K_i(x_0) \widehat{\theta}(x, \widehat{V}_i) / \sum_{i=1}^{n} T_i K_i(x_0),$$

for $K_i(x_0) = 1(X_i = x_0)$ when X is discrete or $K_i(x_0) = k_h(X_i - x_0)$ when X is continuous, where $k_h(u) = k(u/h)/h$, k is a kernel, and h is a bandwidth such as $h \to 0$ as $n \to 0$. Finally, the estimator of the effect (3.6) is

$$\widehat{\theta}_S = \sum_{i=1}^{n} T_i \widehat{\theta}(X_i, \widehat{V}_i) / \sum_{i=1}^{n} T_i.$$

4.4 Step 4: Estimation of counterfactual distributions

Based on equations (3.9) and (3.10), the estimator (or sample analog) of the counterfactual distribution is:

$$\widehat{G}^s_{Y_{\langle t|k,r\rangle}}(y) = \sum_{i=1}^{n} \Lambda(\widehat{W}_i^{\mathrm{T}} \widehat{\beta}_t(y)) 1[\widehat{V}_i > \Lambda(R_i^{\mathrm{T}} \widehat{\pi}_r(0))] / n^s_{kr},$$

where the average is taken over the sample values of $\widehat{V}_i$ and Z_i in group k, $n^s_{kr} = \sum_{i=1}^{n} 1[\widehat{V}_i > \Lambda(R_i^{\mathrm{T}} \widehat{\pi}_r(0))]$, $\widehat{\beta}_t(y)$ is the distribution regression estimator of step 2 in

[3] We can use the generalized inverse

$$\widehat{q}_S(\tau, x) = \int_0^{\infty} 1(\widehat{G}_S(y, x) \leq \tau) dy - \int_{-\infty}^{0} 1(\widehat{G}_S(y, x) > \tau) dy,$$

which does not require that the estimator of the DSF $y \mapsto \widehat{G}_S(y, x)$ be monotone.

group t, and $\widehat{\pi}_r(0)$ is the distribution regression estimator of step 1 in group r. Here we are estimating the components $F^s_{Y_t}$ by logistic distribution regression in group t and the component $F^s_{Z_k}$ by the empirical distribution in group k.

4.5 Inference

We use weighted bootstrap to make inference on all the objects of interest (Praestgaard and Wellner, 1993; Hahn, 1995). This method obtains the bootstrap version of the estimator of interest by repeating all the estimation steps including random draws from a distribution as sampling weights. The weights should be positive and come from a distribution with unit mean and variance such as the standard exponential. Weighted bootstrap has some theoretical and practical advantages over empirical bootstrap. Thus, it is appealing that the consistency can be proven following the strategy set forth by Ma and Kosorok (2005), and the smoothness induced by the weights helps dealing with discrete covariates with small cell sizes. The implementation of the bootstrap for the local and global effects is summarized in the following algorithm:

Algorithm 4 (Weighted Bootstrap) *For $b = 1, \ldots, B$, repeat the following steps: (1) Draw a set of weights $(\omega_1^b, \ldots, \omega_n^b)$ i.i.d. from a distribution that satisfies Condition 1(b) such as the standard exponential distribution. (2) Obtain the bootstrap draws of the control function, $\widehat{V}_i^b = \Lambda(R_i^{\mathrm{T}}\widehat{\pi}^b(C_i))$, $i = 1, \ldots, n$, where for $c \in \mathcal{C}_n$,*

$$\widehat{\pi}^b(c) = \arg\max_{\pi \in \mathbb{R}^{d_r}} \sum_{i=1}^n \omega_i^b \left[1\{C_i \le c\} \log \Lambda(R_i^{\mathrm{T}}\pi)) + 1\{C_i > c\} \log \Lambda(-R_i^{\mathrm{T}}\pi)\right].$$

(3) Obtain the bootstrap draw of the local effect, $\widehat{\theta}^b(x, v)$. For the LASF, $\widehat{\theta}^b(x, v) = \widehat{\mu}^b(x, v) = w(x, v)^{\mathrm{T}}\widehat{\beta}^b$, where

$$\widehat{\beta}^b = \left[\sum_{i=1}^n \omega_i^b \widehat{W}_i^b (\widehat{W}_i^b)^{\mathrm{T}} T_i\right]^{-1} \sum_{i=1}^n \omega_i^b \widehat{W}_i^b Y_i T_i, \quad \widehat{W}_i^b := w(X_i, \widehat{V}_i^b).$$

For the LDSF, $\widehat{\theta}^b(x, v) = \widehat{G}^b(y, x, v) = \Lambda(w(x, v)^{\mathrm{T}}\widehat{\beta}^b(y))$, where

$$\widehat{\beta}^b(y) = \arg\max_{b \in \mathbb{R}^{d_w}} \sum_{i=1}^n \omega_i^b \left[1\{Y_i \le y\} \log \Lambda(b^{\mathrm{T}}\widehat{W}_i^b) + 1\{Y_i > y\} \log \Lambda(-b^{\mathrm{T}}\widehat{W}_i^b)\right] T_i.$$

For the LQSF, $\widehat{\theta}^b(x,v) = \widehat{q}^b(\tau,x,v) = w(x,v)^{\mathrm{T}}\widehat{\beta}^b(\tau)$*, where*

$$\widehat{\beta}^b(\tau) = \arg\min_{b\in\mathbb{R}^{d_w}} \sum_{i=1}^n \omega_i^b \rho_\tau(Y_i - b^{\mathrm{T}}\widehat{W}_i^b)T_i.$$

(4) Obtain the bootstrap draw of the global effects as

$$\widehat{\theta}_S^b(x) = \sum_{i=1}^n \omega_i^b T_i \widehat{\theta}^b(x,\widehat{V}_i^b) / \sum_{i=1}^n \omega_i^b T_i,$$

$$\widehat{\theta}_S^b(x \mid x_0) = \sum_{i=1}^n \omega_i^b T_i K_i(x_0) \widehat{\theta}^b(x,\widehat{V}_i^b) / \sum_{i=1}^n \omega_i^b T_i K_i(x_0),$$

or

$$\widehat{\theta}_S^b = \sum_{i=1}^n \omega_i^b T_i \widehat{\theta}^b(X_i,\widehat{V}_i^b) / \sum_{i=1}^n \omega_i^b T_i.$$

4.6 Asymptotic theory

We derive large sample theory for some of the local and global effects. We focus on average effects for the sake of brevity. The theory for distribution and quantile effects can be derived using similar arguments, see, for example, Chernozhukov et al. (2015) and Chernozhukov et al. (2017). Through the analysis we treat the dimensions of the flexible specifications used in all the steps as fixed, so that the model parameters are estimable at a $\sqrt{n}$ rate. The model is still semiparametric because some of the parameters are function-valued such as the parameters of the control variable.[4]

In what follows, we shall use the following notation. We let the random vector $A = (Y * 1(C > 0), C, Z, V)$ live on some probability space $(\Omega_0, \mathcal{F}_0, P)$. Thus, the probability measure P determines the law of A or any of its elements. We also let $A_1, ..., A_n$, i.i.d. copies of A, live on the complete probability space $(\Omega, \mathcal{F}, \mathbb{P})$, which contains the infinite product of $(\Omega_0, \mathcal{F}_0, P)$. Moreover, this probability space can be suitably enriched to carry also the random weights that appear in the weighted bootstrap. The distinction between the two laws P and $\mathbb{P}$ is helpful to simplify the notation in the proofs and in the analysis. Calligraphic letters such as $\mathcal{Y}$ and $\mathcal{X}$

[4]Chernozhukov at al. (2017) discuss the trade-offs between imposing parametric restrictions in the model and the support conditions required for nonparametric identification of the effects of interest.

denote the supports of $Y * 1(C > 0)$ and X; and $\mathcal{YX}$ denotes the joint support of (Y, X). Unless explicitly mentioned, all functions appearing in the statements are assumed to be measurable.

We now state formally the assumptions. The first assumption is about sampling and the bootstrap weights.

Condition 1 (Sampling and Bootstrap Weights) *(a) Sampling: the data* $\{Y_i * 1(C_i > 0), C_i, Z_i\}_{i=1}^n$ *are a sample of size* n *of independent and identically distributed observations from the random vector* $(Y * 1(C > 0), C, Z)$*. (b) Bootstrap weights:* $(\omega_1, ..., \omega_n)$ *are i.i.d. draws from a random variable* $\omega \geq 0$*, with* $\mathbb{E}_P[\omega] = 1$*,* $\mathrm{Var}_P[\omega] = 1$*, and* $\mathbb{E}_P|\omega|^{2+\delta} < \infty$ *for some* $\delta > 0$*; live on the probability space* $(\Omega, \mathcal{F}, \mathbb{P})$*; and are independent of the data* $\{Y_i * 1(C_i > 0), C_i, Z_i\}_{i=1}^n$ *for all* n*.*

The second assumption is about the first stage where we estimate the control function

$$\vartheta_0(c, z) := F_C(c \mid z).$$

We assume a logistic distribution regression model for the conditional distribution of C in the trimmed support, $\overline{\mathcal{C}}$, that excludes censored and extreme values of C. The purpose of the upper trimming is to avoid the upper tail in the modeling and estimation of the control variable, and to make the eigenvalue assumption in Condition 2(b) more plausible. We consider a fixed trimming rule, which greatly simplifies the derivation of the asymptotic properties. Throughout this section, we use bars to denote trimmed supports with respect to C, e.g., $\overline{\mathcal{CZ}} = \{(c, z) \in \mathcal{CZ} : c \in \overline{\mathcal{C}}\}$, and $\overline{\mathcal{V}} = \{\vartheta_0(c, z) : (c, z) \in \overline{\mathcal{CZ}}\}$.

Condition 2 (First Stage) *(a) Trimming: we consider the trimming rule as defined by the indicator* $T = \mathbf{1}(C \in \overline{\mathcal{C}})$*. (b) Model: the distribution of* C *conditional on* Z *follows the distribution regression model in the trimmed support* $\overline{\mathcal{C}}$*,* i.e.*,*

$$F_C(c \mid Z) = F_C(c \mid R) = \Lambda(R^{\mathrm{T}}\pi_0(c)), \quad R = r(Z),$$

for all $c \in \overline{\mathcal{C}}$*, where* Λ *is the logit link function; the coefficients* $c \mapsto \pi_0(c)$ *are three times continuously differentiable with uniformly bounded derivatives;* $\overline{\mathcal{R}}$ *is compact;*

and the minimum eigenvalue of $\mathbb{E}_P\left[\Lambda(R^{\mathrm{T}}\pi_0(c))[1-\Lambda(R^{\mathrm{T}}\pi_0(c))]RR^{\mathrm{T}}\right]$ *is bounded away from zero uniformly over* $c \in \overline{\mathcal{C}}$.

For $c \in \overline{\mathcal{C}}$, let

$$\widehat{\pi}^b(c) \in \arg\min_{\pi\in\mathbb{R}^{\dim(R)}} \sum_{i=1}^{n} \omega_i\{1(C_i \le c)\log\Lambda(R_i^{\mathrm{T}}\pi) + 1(C_i > c)\log\Lambda(-R_i^{\mathrm{T}}\pi)\},$$

where either $\omega_i = 1$ for the unweighted sample, to obtain the estimator; or ω_i are the bootstrap weights to obtain bootstrap draws of the estimator. Then set

$$\vartheta_0(c,r) = \Lambda(r^{\mathrm{T}}\pi_0(c));\ \widehat{\vartheta}^b(c,r) = \Lambda(r^{\mathrm{T}}\widehat{\pi}^b(c)),$$

if $(c,r) \in \overline{\mathcal{CR}}$, and $\vartheta_0(c,r) = \widehat{\vartheta}^b(c,r) = 0$ otherwise.

Theorem 4 of Chernozhukov et al. (2015) established the asymptotic properties of the DR estimator of the control function. We repeat the result here as a lemma for completeness and to introduce notation that will be used in the results below. Let $\|f\|_{T,\infty} := \sup_{a\in\mathcal{A}}|T(c)f(a)|$ for any function $f : \mathcal{A} \mapsto \mathbb{R}$, and $\lambda = \Lambda(1-\Lambda)$, the density of the logistic distribution.

Lemma 2 (First Stage) *Suppose that Conditions 1 and 2 hold. Then, (1)*

$$\begin{aligned}
\sqrt{n}(\widehat{\vartheta}^b(c,r) - \vartheta_0(c,r)) &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n} e_i\ell(A_i,c,r) + o_{\mathbb{P}}(1) \rightsquigarrow \Delta^b(c,r) \text{ in } \ell^\infty(\overline{\mathcal{CR}}),\\
\ell(A,c,r) &:= \lambda(r^{\mathrm{T}}\pi_0(c))[1\{C\le c\} - \Lambda(R^{\mathrm{T}}\pi_0(c))] \times\\
&\quad \times r^{\mathrm{T}}\mathbb{E}_P\left\{\Lambda(R^{\mathrm{T}}\pi_0(c))[1-\Lambda(R^{\mathrm{T}}\pi_0(c))]RR^{\mathrm{T}}\right\}^{-1}R,\\
\mathbb{E}_P[\ell(A,c,r)] &= 0, \mathbb{E}_P[T\ell(A,C,R)^2] < \infty,
\end{aligned}$$

where $(c,r) \mapsto \Delta^b(c,r)$ *is a Gaussian process with uniformly continuous sample paths and covariance function given by* $\mathbb{E}_P[\ell(A,c,r)\ell(A,\tilde{c},\tilde{r})^{\mathrm{T}}]$. *(2) There exists* $\widetilde{\vartheta}^b : \overline{\mathcal{CR}} \mapsto [0,1]$ *that obeys the same first order representation uniformly over* $\overline{\mathcal{CR}}$, *is close to* $\widehat{\vartheta}^b$ *in the sense that* $\|\widetilde{\vartheta}^b - \widehat{\vartheta}^b\|_{T,\infty} = o_{\mathbb{P}}(1/\sqrt{n})$ *and, with probability approaching one, belongs to a bounded function class* Υ *such that the covering entropy*

satisfies:[5]

$$\log N(\epsilon, \Upsilon, \|\cdot\|_{T,\infty}) \lesssim \epsilon^{-1/2}, \quad 0 < \epsilon < 1.$$

The next assumptions are about the second stage. We assume a flexible linear model for the conditional distribution of Y given (X, V) in the trimmed support $C \in \overline{\mathcal{C}}$, impose compactness conditions, and provide sufficient conditions for identification of the parameters. Compactness is imposed over the trimmed support and can be relaxed at the cost of more complicated and cumbersome proofs.

Condition 3 (Second Stage) *(a) Model: the expectation of Y conditional on (X, V) in the trimmed support $C \in \overline{\mathcal{C}}$ is*

$$\mathbb{E}(Y \mid X, V, C \in \overline{\mathcal{C}}) = W^{\mathrm{T}}\beta_0, \quad V = F_{C|Z}(C \mid Z), \quad W = w(X, V).$$

(b) Compactness and moments: the set $\overline{\mathcal{W}}$ is compact; the derivative vector $\partial_v w(x, v)$ exists and its components are uniformly continuous in $v \in \overline{\mathcal{V}}$, uniformly in $x \in \overline{\mathcal{X}}$, and are bounded in absolute value by a constant, uniformly in $(x, v) \in \overline{\mathcal{X}\mathcal{V}}$; $\mathbb{E}(Y^2 \mid C \in \overline{\mathcal{C}}) < \infty$; and $\beta_0 \in \mathcal{B}$, where $\mathcal{B}$ is a compact subset of $\mathbb{R}^{d_w}$. (c) Identification and nondegeneracy: the matrix $J := \mathbb{E}_P[WW^{\mathrm{T}}\ T]$ is of full rank; and the matrix $\Omega := \mathrm{Var}_P[f_1(A) + f_2(A)]$ is finite and is of full rank, where

$$f_1(A) := \{W^{\mathrm{T}}\beta_0 - Y\}WT,$$

and, for $\dot{W} = \partial_v w(X, v)|_{v=V}$,

$$f_2(A) := \mathbb{E}_P[\{[W^{\mathrm{T}}\beta_0 - Y]\dot{W} + W^{\mathrm{T}}\beta_0 W\}T\ell(a, C, Z)]\big|_{a=A}.$$

Let

$$\widehat{\beta} = \arg\min_{\beta \in \mathbb{R}^{\dim(W)}} \sum_{i=1}^{n} T_i(Y_i - \beta^{\mathrm{T}}\widehat{W}_i)^2, \quad \widehat{W}_i = w(X_i, \widehat{V}_i), \quad \widehat{V}_i = \widehat{\vartheta}(C_i, R_i),$$

[5] See Appendix B for a definition of the covering entropy.

where $\widehat{\vartheta}$ is the estimator of the control function in the unweighted sample; and

$$\widehat{\beta}^b = \arg\min_{\beta\in\mathbb{R}^{\dim(W)}} \sum_{i=1}^n \omega_i T_i (Y_i - \beta^{\mathrm{T}}\widehat{W}_i^b)^2, \quad \widehat{W}_i^b = w(X_i, \widehat{V}_i^b), \quad \widehat{V}_i^b = \widehat{\vartheta}^b(C_i, R_i),$$

where $\widehat{\vartheta}^b$ is the estimator of the control function in the weighted sample. The following lemma establishes a central limit theorem and a central limit theorem for the bootstrap for the estimator of the coefficients in the second stage.

Let $\rightsquigarrow_{\mathbb{P}}$ denote bootstrap consistency, *i.e.* weak convergence conditional on the data in probability as defined in Appendix B.1.

Lemma 3 (CLT and Bootstrap FCLT for $\widehat{\beta}$) *Under Conditions 1–3, in $\mathbb{R}^{d_w}$,*

$$\sqrt{n}(\widehat{\beta} - \beta_0) \rightsquigarrow J^{-1}G, \quad \textit{and} \quad \sqrt{n}(\widehat{\beta}^b - \widehat{\beta}) \rightsquigarrow_{\mathbb{P}} J^{-1}G,$$

where $G \sim N(0, \Omega)$ and J and Ω are defined in Assumption 3(c).

The properties of the estimator of the LASF, $\widehat{\mu}(x, v) = w(x, v)^{\mathrm{T}}\widehat{\beta}$, and its bootstrap version, $\widehat{\mu}^b(x, v) = w(x, v)^{\mathrm{T}}\widehat{\beta}^b$, are a corollary of Lemma 3.

Corollary 2 (FCLT and Bootstrap FCLT for LASF) *Under Assumptions 1–3, in $\ell(\overline{\mathcal{XV}})$,*

$$\sqrt{n}(\widehat{\mu}(x, v) - \mu(x, v)) \rightsquigarrow Z(x, v) \textit{ and } \sqrt{n}(\widehat{\mu}^b(x, v) - \widehat{\mu}(x, v)) \rightsquigarrow_{\mathbb{P}} Z(x, v),$$

where $(x, v) \mapsto Z(x, v) := w(x, v)^{\mathrm{T}} J^{-1} G$ is a zero-mean Gaussian process with covariance function

$$\mathrm{Cov}_P[Z(x_0, v_0), Z(x_1, v_1)] = w(x_0, v_0)^{\mathrm{T}} J^{-1} \Omega J^{-1} w(x_1, v_1).$$

To obtain the properties of the estimator of the ASFs, we define $W_x := w(x, V)$, $\widehat{W}_x := w(x, \widehat{V})$, and $\widehat{W}_x^b := w(x, \widehat{V}^b)$. The estimator and its bootstrap draw of the ASF in the trimmed support, $\mu_S(x) = \mathbb{E}_P\{\beta_0^{\mathrm{T}} W_x \mid T = 1\}$, are $\widehat{\mu}_S(x) = \sum_{i=1}^n \widehat{\beta}^{\mathrm{T}} \widehat{W}_{xi} T_i / n_T$, and $\widehat{\mu}_S^b(x) = \sum_{i=1}^n e_i \widehat{\beta}^{bT} \widehat{W}_{xi}^b T_i / n_T^b$, where $n_T = \sum_{i=1}^n T_i$ and $n_T^b = \sum_{i=1}^n e_i T_i$. The estimator and its bootstrap draw of the ASF on the treated in the trimmed support, $\mu_S(x \mid x_0) = \mathbb{E}_P\{\beta_0^{\mathrm{T}} W_x \mid T = 1, X = x_0\}$, are $\widehat{\mu}_S(x \mid x_0) =$

$\sum_{i=1}^n \widehat{\beta}^{\mathrm{T}}\widehat{W}_{xi}K_i(x_0)T_i/n_T(x_0)$, and $\widehat{\mu}_S^b(x) = \sum_{i=1}^n e_i\widehat{\beta}^{bT}\widehat{W}_{xi}^b\ K_i(x_0)T_i/n_T^b(x_0)$, where $n_T(x_0) = \sum_{i=1}^n K_i(x_0)T_i$ and $n_T^b(x_0) = \sum_{i=1}^n e_iK_i(x_0)T_i$. Let $p_T := P(T = 1)$ and $p_T(x) := P(T = 1, X = x)$. The next result gives large sample theory for these estimators. The theory for the ASF on the treated is derived for X discrete, which is the relevant case in our empirical application.

Theorem 5 (FCLT and Bootstrap FCLT for ASF) *Under Assumptions 1–3, in $\ell(\overline{\mathcal{X}})$,*

$$\sqrt{np_T}(\widehat{\mu}_S(x) - \mu_S(x)) \rightsquigarrow Z(x) \text{ and } \sqrt{np_T}(\widehat{\mu}_S^b(x) - \widehat{\mu}_S(x)) \rightsquigarrow_{\mathbb{P}} Z(x),$$

where $x \mapsto Z(x)$ is a zero-mean Gaussian process with covariance function

$$\mathrm{Cov}_P[Z(x_0), Z(x_1)] = \mathrm{Cov}_P[W_{x_0}^{\mathrm{T}}\beta_0 + \sigma_{x_0}(A), W_{x_1}^{\mathrm{T}}\beta_0(v) + \sigma_{x_1}(A) \mid T = 1],$$

with

$$\sigma_x(A) = \mathbb{E}_P\{W_x^{\mathrm{T}}T\}J^{-1}[f_1(A) + f_2(A)] + \mathbb{E}_P\{\dot{W}_x^{\mathrm{T}}\beta_0 T\ell(a, X, R)\}\big|_{a=A}.$$

Also, if $p_T(x_0) > 0$, in $\ell(\overline{\mathcal{X}})$,

$$\sqrt{np_T(x_0)}(\widehat{\mu}_S(x \mid x_0) - \mu_S(x \mid x_0)) \rightsquigarrow Z(x \mid x_0) \text{ and}$$
$$\sqrt{np_T(x_0)}(\widehat{\mu}_S^b(x \mid x_0) - \widehat{\mu}_S(x \mid x_0)) \rightsquigarrow_{\mathbb{P}} Z(x \mid x_0),$$

where $x \mapsto Z(x \mid x_0)$ is a zero-mean Gaussian process with covariance function

$$\mathrm{Cov}_P[Z(x \mid x_0), Z(\tilde{x} \mid x_0)] = \mathrm{Cov}_P[W_x^{\mathrm{T}}\beta_0 + \psi_x(A), W_{\tilde{x}}^{\mathrm{T}}\beta_0 + \psi_{\tilde{x}}(A) \mid T = 1, X = x_0],$$

Theorem 5 can be used to construct confidence bands for the ASFs, $x \mapsto \mu_S(x)$ and $x \mapsto \mu_S(x \mid x_0)$, over regions of values of x via Kolmogorov-Smirnov type statistics and weighted bootstrap, and to construct confidence intervals for average treatment effects, $\mu(x_1) - \mu(x_0)$ and $\mu(x_1 \mid x_0) - \mu(x_0 \mid x_0)$, via t-statistics and weighted bootstrap.

5 Application: United Kingdom wage regressions

We now investigate two important issues related to the wage level of female workers in the United Kingdom and the rate of their wage growth. First, we examine the impact of selection bias from the hours decision in estimating the returns to human capital. Second, we provide a decomposition of earnings growth which includes a contribution resulting from selection bias. We use data from the United Kingdom Family Expenditure Survey (FES) for the years 1978 to 1999. Blundell et al. (2003) study male wage growth and Blundell et al. (2007) examine wage inequality for both males and females using the same data source. We employ the same data selection rules and refer the reader to these earlier papers for details. The FES is a repeated cross section of households and contains detailed information on the number of weekly hours worked and the hourly wage of the individual. We restrict the data to those who report an education level and only include working women who report working weekly hours of 70 or less and an hourly wage of at least 0.01 pounds. This reduces the total number of observations from 96,402 to 94,985. This produces a data set of over 4,100 observations per year and with approximately 2,600 working females.

The outcome variable is the log-hourly wage defined as the nominal weekly earnings divided by the number of hours worked and deflated by the quarterly UK retail price index. Following Blundell et al. (2003), we use the simulated out-of-work benefits income as an exclusion restriction in the hours equation. We refer to their paper for details and note that the UK benefits system makes this restriction appropriate since, in contrast to other European countries, unemployment benefits are not related to income prior to the period out of work. Blundell et al. (2007) argue that the system of housing benefits may still have a positive relationship with in-work potential. However, we do not consider these additional issues and refer to Blundell et al. (2007) for a potential solution using a monotonicity restriction in place of an exclusion restriction in the hours equation. Equations (2.1) and (2.2) characterize the model of Blundell et al. (2003) when g and h are linear and separable and ε and η are normally distributed.

Figure 1A reports the female participation rate over the sample period. Figure 1B reports the average number of hours for all females and those reporting positive

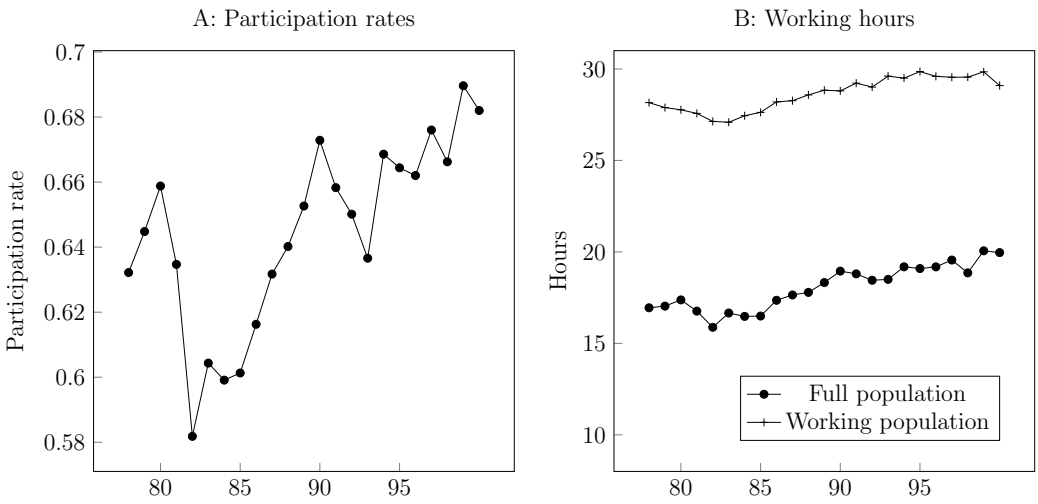

Figure 1: Descriptive statistics of the data set

hours respectively. Recall that our control function exploits the variation in both the extensive and intensive margins of the hours decisions. Figure 1A illustrates that participation was around 65 percent in the years before the recession in the beginning 1980's. Participation drops to a sample period low of 58 percent in 1982, but subsequently increases and almost reaches 70 percent at the end of the sample period. The figures for the average hours show similar trends but most notably there is significant variation in average hours over time for the sample of workers. The figures illustrate the utility of exploiting the number of hours rather than just the binary outcome if they are available.

We use the following variables for our empirical analysis. We use three different education levels; (1) a dummy variable indicating the individual left school at the age of 16 years or younger, (2) a dummy variable for left school at the age of 17 or 18 years, and (3) a dummy variable for left school at the age of 19 years or older. We use age and age squared and interact these with the level of education. In addition, we use a dummy variable indicating that the individual lives together with a partner and we use 12 dummy variables indicating the region in the UK in which the individual lives. We pool the data for four consecutive years, *i.e.* 1978-81, 1982-1985, 1986-1989, 1990-1993, 1994-1997 and 1998-2000 noting that the last period is only 3 years.

	Mean	Q1	Q2	Q3
Leaving school at the age of 17-18				
1978-1981	0.259	0.149	0.252	0.349
	(0.227,0.290)	(0.221,0.283)	(0.221,0.283)	(0.314,0.384)
1986-1989	0.275	0.195	0.292	0.335
	(0.248,0.302)	(0.164,0.227)	(0.265,0.320)	(0.294,0.376)
1998-2000	0.273	0.223	0.299	0.335
	(0.245,0.301)	(0.267,0.331)	(0.124,0.377)	(0.301,0.370)
Leaving school at the age of 19 or older				
1978-1981	0.658	0.564	0.746	0.809
	(0.624,0.692)	(0.511,0.618)	(0.704,0.787)	(0.771,0.849)
1986-1989	0.579	0.549	0.701	0.691
	(0.551,0.608)	(0.512,0.585)	(0.671,0.732)	(0.660,0.722)
1998-2000	0.597	0.521	0.701	0.703
	(0.567,0.628)	(0.473, 0.569)	(0.665,0.736)	(0.669,0.737)

Table 1: Estimates of the returns to education without correction for sample selection. Bootstrapped confidence intervals are in between parentheses.

5.1 Returns to human capital

Given the changes in working hours over the sample period, we investigate the impact of selection on the return to human capital. We first examine the returns to schooling. Table 1 reports the impact of education on wages estimated by quantile regression unadjusted for selection. The reported results in the first column are the absolute values of the average treatment effects of the difference between the lowest level of education and any higher level of education. Similarly, columns 2 to 4 report the absolute values of the quantile treatment effects.[6] The results in Table 1 indicate that there is generally a larger coefficient at higher quantiles. There is also evidence that there is an increase in the return to education over time at some quantiles.

Table 2 reports the results of the local average and quantile treatment effects. We report the absolute values of these effects based on the subsample of individuals with the lowest level of education. We account for sample selection by including V and V^2 as well as interaction terms of V with all the regressors discussed above. Note that we report our results for values of V at the median and higher as it appears that our identification requirements are not satisfied at lower quantiles.

[6]We calculate the average treatment effects for the medium education level as the difference in the average wage among the lowest educated and $\sum_{educ=\text{“}low''} P(x, educ = \text{“}medium'')\widehat{\beta}$, where P is a polynomial. We use distribution regression for the quantile treatment effect to estimate the distribution and calculate the quantile of that distribution.

	Mean	Q1	Q2	Q3
		$V = 0.5$		
1978-1981	0.263	0.151	0.217	0.300
	(0.200,0.327)	(0.074,0.229)	(0.139,0.296)	(0.210,0.390)
1986-1989	0.309	0.122	0.304	0.426
	(0.252,0.366)	(0.059,0.0186)	(0.254,0.354)	(0.363,0.488)
1998-2000	0.281	0.178	0.318	0.411
	(0.232,0.330)	(0.116,0.241)	(0.261,0.375)	(0.341,0.481)
		$V = 0.75$		
1978-1981	0.258	0.154	0.256	0.346
	(0.225,0.291)	(0.115,0.194)	(0.224,0.287)	(0.307,0.385)
1986-1989	0.279	0.199	0.307	0.350
	(0.253,0.305)	(0.168,0.230)	(0.277,0.336)	(0.311,0.389)
1998-2000	0.278	0.235	0.306	0.343
	(0.250,0.306)	(0.196,0.276)	(0.273,0.337)	(0.310,0.377)
Leaving school at the age of 19 or older				
		$V = 0.5$		
1978-1981	0.742	0.610	0.911	0.954
	(0.679,0.805)	(0.465,0.755)	(0.828,0.994)	(0.880,1.028)
1986-1989	0.638	0.456	0.814	0.877
	(0.579,0.697)	(0.329,0.583)	(0.744,0.883)	(0.829,0.926)
1998-2000	0.653	0.385	0.779	0.889
	(0.592,0.713)	(0.314,0.456)	(0.688,0.870)	(0.816,0.963)
		$V = 0.75$		
1978-1981	0.661	0.564	0.747	0.802
	(0.627,0.696)	(0.512,0.615)	(0.706,0.788)	(0.758,0.846)
1986-1989	0.589	0.562	0.717	0.705
	(0.561,0.616)	(0.528,0.596)	(0.686,0.747)	(0.674,0.736)
1998-2000	0.601	0.531	0.706	0.712
	(0.570,0.601)	(0.483,0.579)	(0.670,0.741)	(0.677,0.748)

Table 2: Estimates of the returns to education, our method using a control function. Bootstrapped confidence intervals are in between parentheses.

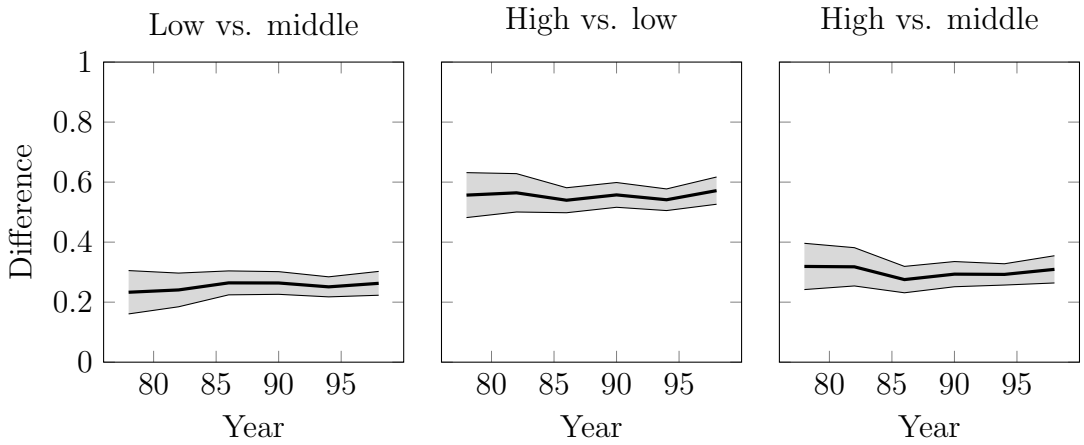

Figure 2A: Global estimates of the average impact of education for the low and middle educated in the selected population in case that they have any other education level.

An examination of Table 2 reveals that the impact of education varies by quantile and by the value of V at which it is evaluated. Looking at the results at the mean, there appears to be some variation in the returns to education for different values of V, but the evidence is not strong statistically. This, in addition to the similarity of these results to the unadjusted results, may suggest that there are no clear indications of selection bias for these quantiles at higher levels of V.

We further explore the role of education by deriving the average and the quantile impact of obtaining a higher education for some qualified groups. These are shown in Figures 2A and 2B. The estimates are based on pooling the data in the same manner as above. The labels "Low", "Middle" and "High" capture the three education groups. Hence, the figure "Low versus middle" in Figure 2A displays the average increase in wages when women of the lowest education group have an education level equal to the middle education level. The figure displayed for $\tau = 0.25$ in Figure 2B looks at this increase at the first quartile of the distribution. The magnitude of the average impact of education for the various educational comparisons is consistent with the estimates in the tables discussed above and the plot over time appears to reveal some cyclical behavior.

We also explore how the return to experience has varied by education group over the sample period by estimating the average derivative with respect to age. Figure 3 presents the derivative for different education levels. The figures represent the

age weighted average derivative based on a weighted average over the sample. The figures show that there is a drastic increase to the return to experience during the 1990s. They also reveal that there is a drastic difference in the rate of wage growth across education groups. Figure 4 reports these derivatives evaluated at ages 25, 40 and 55 years and these represent the local average responses. There is a strong positive relationship between wage growth and age at 25 years and the effect is particularly strong for the highest educated. Moreover, the effect increases notably over the sample period with large increases in the 1990s. The effect is notably lower although still positive at the age of 40 years. The differences by education groups are less dramatic. At 55 years, wages do not appear to be generally increasing with age. In fact, there appears to be evidence that the real wage is decreasing for the highest education group.

5.2 Decomposition of the wage increase

The above evidence regarding the impact on human capital and the role of selection on wages suggest each has played a role in the evolution of wages for working females in our sample period. We investigate their respective contributions by following Blundell et al. (2003) to decompose female wage growth into the selection component, composition component and structure component as introduced in Section 3.[7] For these components, we set 1982 as the base year (*i.e.* year 0 as in (3.11)). This choice is based on Figure 1. We do not focus on these components, but report the differences in the quantiles. For example, the selection component equals:

$$\Delta_\tau^1 = \mathbb{Q}_\tau(Y_{\langle t|t,t\rangle} \mid C_t > 0) - \mathbb{Q}_\tau(Y_{\langle t|t,0\rangle} \mid C_0 > 0)$$

and similarly, we introduce Δ_τ^2 and Δ_τ^3.

Figure 5 shows the time series of the different components and the total difference in wages from 1982 to 1999. Similar to Blundell et al. (2003), who find a large change in the wage dispersion for males in this period, we find the total increase to be much larger for the second (Q2) and third quartiles (Q3) than for the bottom

[7] Blundell et al. (2003) provide the decomposition of male wages in the same period, while employing a parametric approach to account for selection.

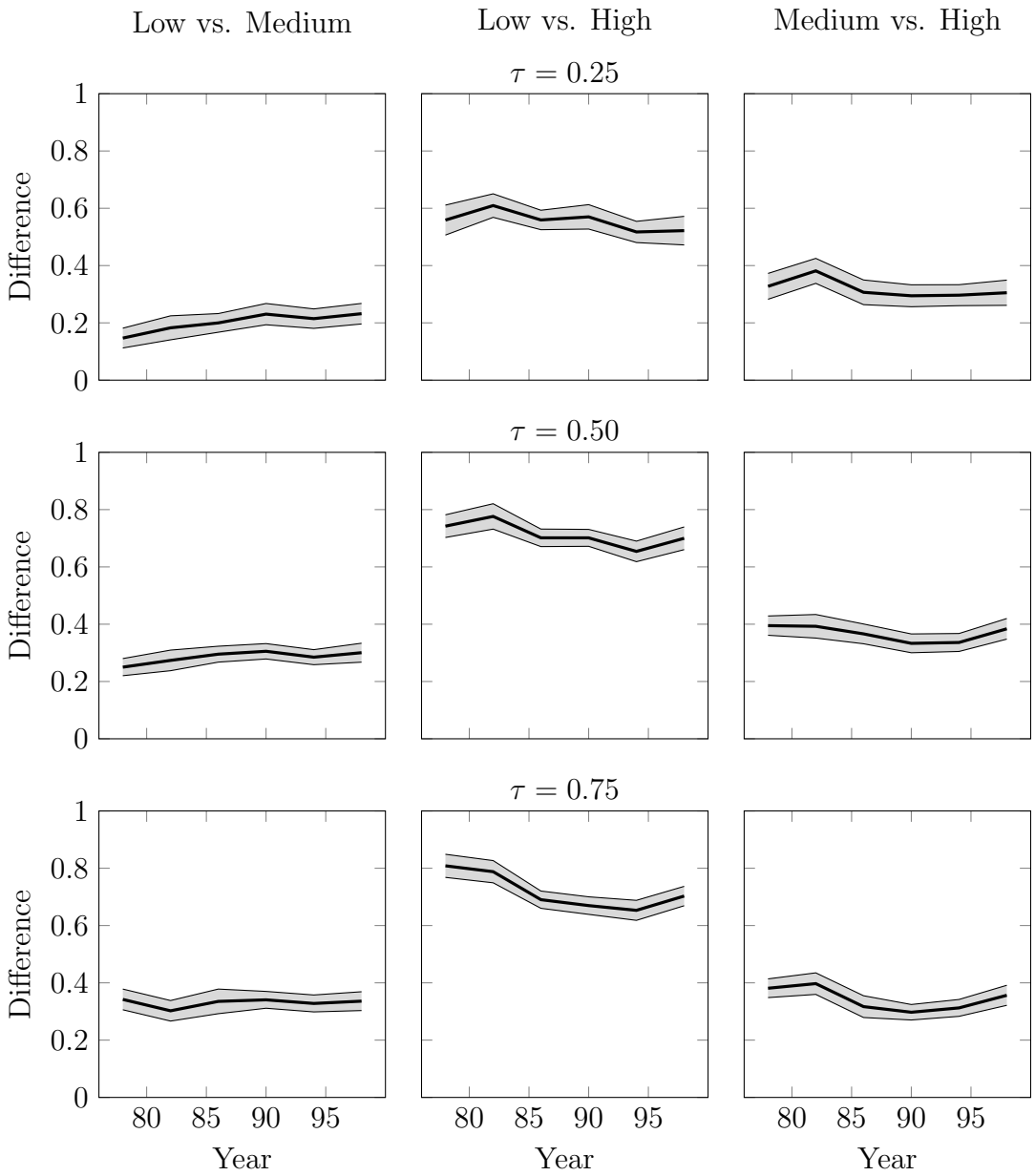

Figure 2B: Global estimates of the quantile impact of education for the low and middle educated in the selected population in case that they have any other education level.

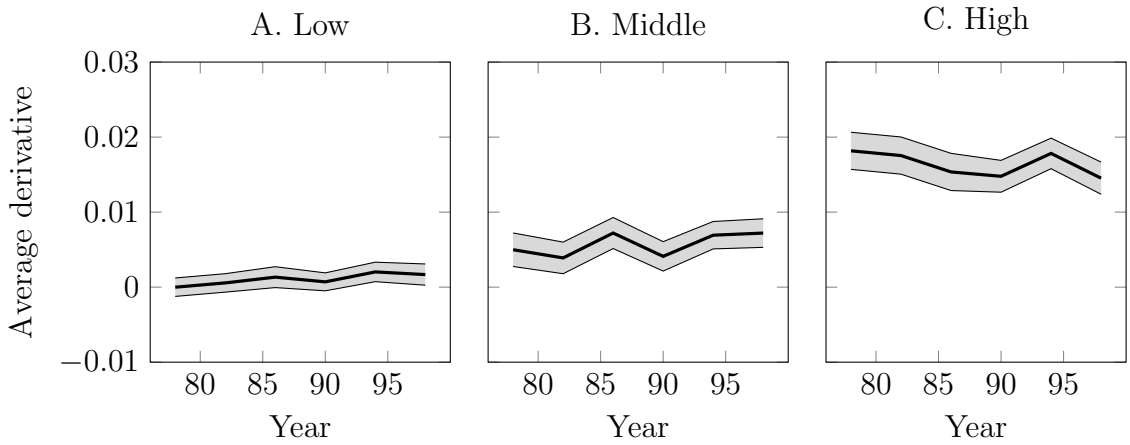

Figure 3: Average derivative of the impact of age on log wages among the selected population.

decile (D1) and first quartile (Q1). More explicitly, while wages grew 26.0 percent for the bottom decile and 31.1 percent for the bottom quartile they grew 42.4 and 48.4 percent at the median and upper quartile respectively. This difference in growth rates is especially drastic since 1991. Notably, we find that the increase in the wages is primarily due to the wage structure component and this is especially true at the bottom of the distribution. There is also evidence that the composition component contributes substantially to wage growth at all quantiles although there is evidence of larger effects at the median and above. The selection component is small in absolute values and negative. The negative effect is expected as comparing later years to 1982 makes the sample more selective. Thus, since it is likely that the "more productive" women were working in 1982, the wages will increase by dropping the less able women in the later years from the sample. The selection effect is largest at D1 and Q1 and almost non-existent at Q3. This is also expected. That is, women at the top of the distribution worked both in 1982 as in any other year and therefore we do not change the composition of the sample, with respect to unobservables, at these higher quantiles by imposing the high 1982 level of selection.

6 Conclusion

This paper examines a nonseparable sample selection model with a selection equation which is based on a partially censored outcome. We account for selection by

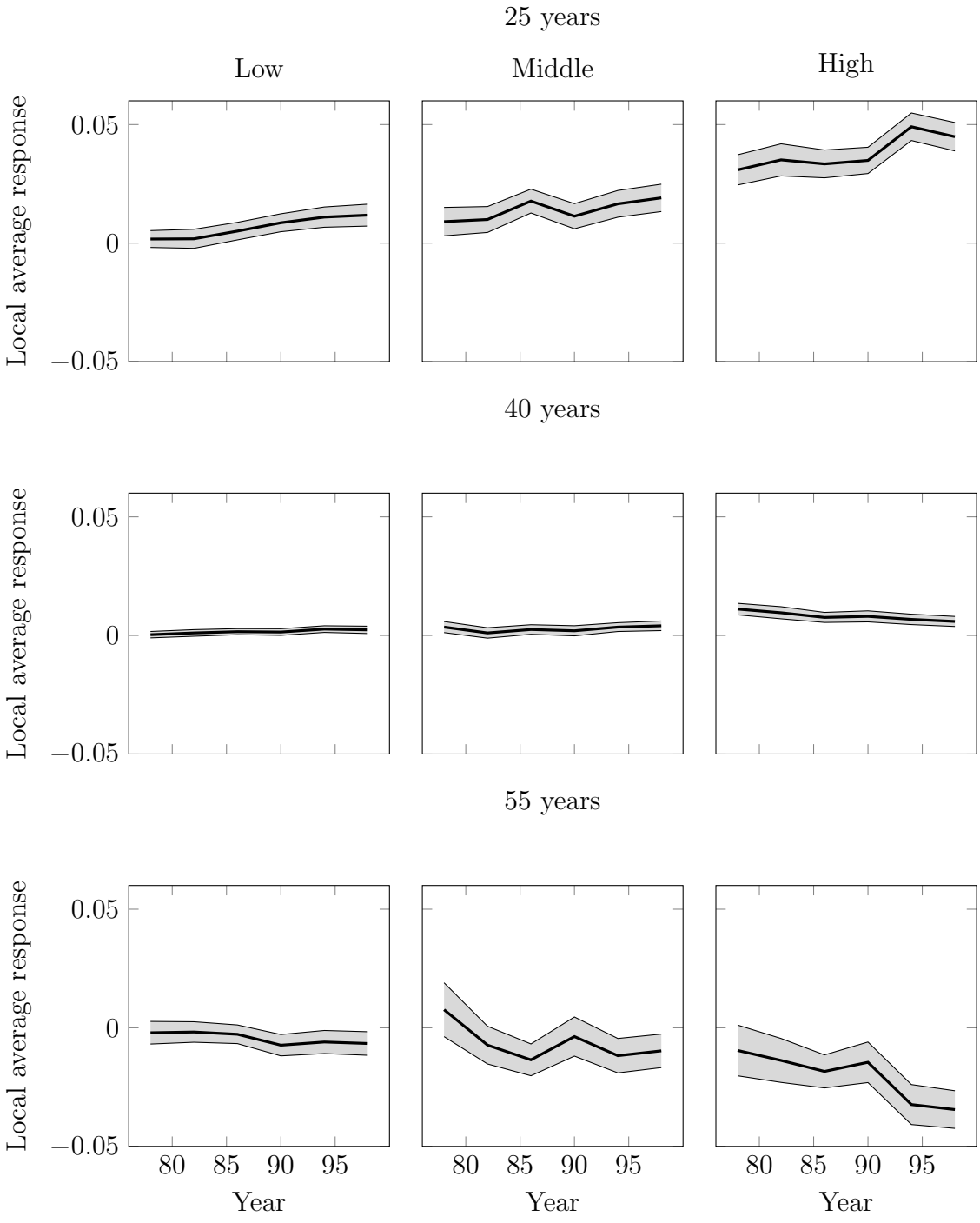

Figure 4: Local average response at different ages among the selected population.

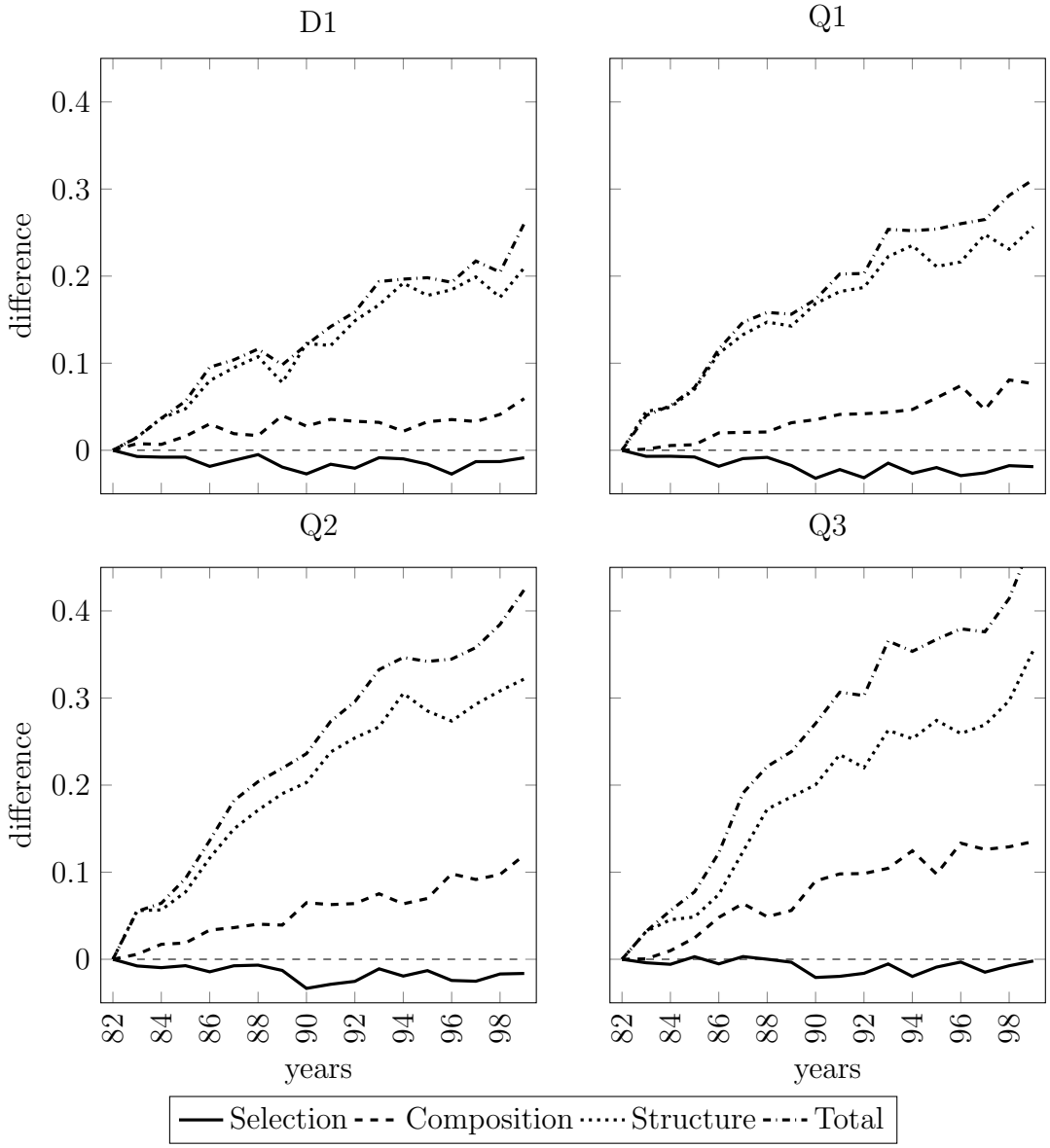

Figure 5: Contributions of the different components of the decomposition of the wage increase since 1982 at different quantiles.

conditioning on an appropriately constructed control function. We show that for this model we are able to identify several economically interesting objects. We categorize these as local effects, which represent estimands conditional on a specific outcome of the control function and global effects, which represent estimands evaluated over a range of values of the control function. For both effects we provide identification results and estimation methods in addition to the related asymptotic theory. We illustrate the utility of our approach in an empirical application focusing on the determinants of wages and wage growth for a sample of United Kingdom females over a period of increasing labor force participation at both the intensive and extensive margins.

References

[1] Amemiya, T. (1978), "The estimation of a simultaneous equation generalized probit model", *Econometrica*, **46**, 1193–1205.

[2] Amemiya, T. (1979), "The estimation of a simultaneous tobit model", *International Economic Review*, **20**, 169–81.

[3] Arellano, M. and S. Bonhomme (2017), "Quantile selection models with an application to understanding changes in wage inequality", *Econometrica*, **85**, 1-28.

[4] Blundell, R., A. Gosling, H. Ichimura and C. Meghir (2007), "Changes in the distribution of male and female wages accounting for employment composition using bounds", *Econometrica*, **75**, 323–63.

[5] Blundell, R., H. Reed and T.M. Stoker (2003), "Interpreting aggregate wage growth: the role of labor market participation", *American Economic Review*, **93**, 1114–31.

[6] Buchinsky, M. (1998), "The dynamics of changes in the female wage distribution in the USA: a quantile regression approach", *Journal of Applied Econometrics*, **13**, 1–30.

[7] Chen, S. (1997), "Semiparametric estimation of the Type-3 Tobit model", *Journal of Econometrics*, **80**, 1–34.

[8] Chernozhukov, V., I. Fernández-Val and A. Kowalski (2015), "Quantile regression with censoring and endogeneity", *Journal of Econometrics*, **186**, 201–21.

[9] Chernozhukov, V., I. Fernández-Val and B. Melly (2013), "Inference on counterfactual distributions", *Econometrica*, **81**, 2205–68.

[10] Chernozhukov, V., I. Fernández-Val, W.K. Newey, S. Stouli and F. Vella (2017), "Semiparametric estimation of structural functions: methods and inference", working paper, University of Bristol.

[11] Chesher, A. (2003), "Identification in nonseparable models", *Econometrica*, **71**, 1401–44.

[12] Dinardo, J., N.M. Fortin and T. Lemieux (1996), "Labor market institutions and the distribution of wages, 1973-1992: a semiparametric approach", *Econometrica*, **64**, 1001–44.

[13] Firpo, S., N.M. Fortin and T. Lemieux (2011), "Decomposition methods in economics" in D. Card and O. Ashenfelter, eds., Handbook of Labor Economics, 4th Edition, Elsevier North Holland, 1–102.

[14] Florens, J., J.J. Heckman, C. Meghir and E. Vytlacil (2008), "Identification of treatment effects using control functions in models with continuous, endogenous treatment and heterogeneous effects", *Econometrica*, **76**, 1191–1206.

[15] Foresi, S. and F. Peracchi (1995), "The conditional distribution of excess returns: an empirical analysis", *Journal of the American Statistical Association*, **90**, 451–66.

[16] Hahn, J. (1995), "Bootstrapping quantile regression estimators", *Econometric Theory*, **11**, 105–21.

[17] Heckman, J.J. (1979), "Sample selection bias as a specification error", *Econometrica*, **47**, 153–61.

[18] Heckman, J.J. and E. Vytlacil (2005), "Structural equations, treatment effects, and econometric policy evaluation", *Econometrica*, **73**, 669–738.

[19] Hoderlein, S. and E. Mammen (2007), "Identification of marginal effects in nonseparable models without monotonicity", *Econometrica*, **75**, 1513–18.

[20] Honoré, B., E. Kyriazidou and C. Udry (1997), "Estimation of type 3 tobit models using symmetric trimming and pairwise comparisons", *Journal of Econometrics*, **76**, 107–28.

[21] Imbens, G.W. and W.K. Newey (2009), "Identification and estimation of triangular equations models without additivity", *Econometrica*, **77**, 1481–1512.

[22] Jun, S.J. (2009), "Local structural quantile effects in a model with a nonseparable control variable", *Journal of Econometrics*, **151**, 82–97.

[23] Koenker, R. and G. Bassett (1978), "Regression quantiles", *Econometrica*, **46**, 33-51.

[24] Lee, M.J. and F. Vella (2006), "A semi-parametric estimator for censored selection models with endogeneity", *Journal of Econometrics*, **130**, 235–52.

[25] Ma, L. and R. Koenker (2006), "Quantile regression methods for recursive structural equation models", *Journal of Econometrics*, **134**, 471–506.

[26] Ma, S. and M. Kosorok (2005), "Robust semiparametric M-estimation and the weighted bootstrap", *Journal of Multivariate Analysis*, **96**, 190-217.

[27] Masten, M. and A. Torgovitsky (2014), "Instrumental variables estimation of a generalized correlated random coefficients model", working paper, Duke University, Durham.

[28] Matzkin, R. (2003), "Nonparametric estimation of nonadditive random functions", *Econometrica*, **71**, 1339–75.

[29] Newey, W.K. (2007), "Nonparametric continuous/discrete choice models", *International Economic Review*, **48**, 1429–39.

[30] Praestgaard, J. and J. Wellner (1993), "Exchangeably weighted bootstraps of the general empirical process", *Annals of Probability*, **21**, 2053–86.

[31] Van der Vaart, A.W. (1998), "*Asymptotic statistics*", Cambridge University Press.

[32] Van der Vaart, A.W. and J.A. Wellner (1996), "*Weak convergence and empirical processes*", Springer, New York.

[33] Vella, F. (1993), "A simple estimator for simultaneous models with censored endogenous regressors", *International Economic Review*, **34**, 441–57.

A Proofs of Section 3

A.1 Lemma 1

Proof. The proof is similar to the proof of Theorem 1 in Newey (2007). For any bounded function $a(\varepsilon)$ and $C > 0$ (and hence $h(Z, \eta) > 0$), by Assumption 1,

$$\mathbb{E}\left[a(\varepsilon) \mid Z = z, \eta = q, C > 0\right] = \mathbb{E}\left[a(\varepsilon) \mid \eta = q, C > 0\right]$$

Since this holds for any function $a(\varepsilon)$ and any $h(Z, \eta) > 0$, Z and ε are independent conditional on η and $C > 0$. The result follows because η is a one-to-one function of $F_{C|Z}(C \mid Z)$ when $C > 0$ since $\eta = h^{-1}(Z, C)$ if $C > 0$ by Assumption 1, and for $c > 0$

$$\begin{aligned} F_{C|Z}(c \mid Z = z) &= \mathbb{P}(\max(h(Z, \eta), 0) \leq c \mid Z = z) \\ &= \mathbb{P}(h(Z, \eta) \leq c \mid Z = z) = \mathbb{P}(\eta \leq h^{-1}(Z, c) \mid Z) = h^{-1}(Z, c), \end{aligned}$$

where we use the normalization $\eta \sim U(0, 1)$. ■

A.2 Theorem 1

Proof. Define the generic local object

$$\theta(x, v) = \mathbb{E}_{\varepsilon}[\Gamma(x, \varepsilon) \mid V = v]$$

for some function $\Gamma(x, e) : \mathcal{X} \times \mathcal{E} \to \mathbb{R}^k; k \in \mathbb{N}^+$, where $\mathcal{E}$ is the support of ε. Using Assumption 1, this equals

$$\theta(x, v) = \mathbb{E}_{\varepsilon}[\Gamma(x, \varepsilon) \mid Z = z, V = v].$$

Since conditional on $Z = z$ and $V = v$, we have that $C = \max\{h(z,v), 0\}$ and since $(x,v) \in \mathcal{XV}$, there is a $z \in Z$ such that $C = h(z,v)$ and hence

$$\begin{aligned}\theta(x,v) &= \mathbb{E}_\varepsilon[\Gamma(x,\varepsilon) \mid Z = z, V = v, C > 0] \\ &= \mathbb{E}_\varepsilon[\Gamma(X,\varepsilon) \mid Z = z, V = v, C > 0],\end{aligned}$$

where the second line is due to $X \subseteq Z$. This second line is identical to the right-hand side of (3.1) when $\Gamma(x,e) = g(x,e)$. Along the same lines, this also proves Corollary 1, with $\Gamma(x,e) = \partial_x g(x,e)$. Note that this also proves identification of the LDSF, since

$$G(y,x,v) = \mathbb{E}_\varepsilon[\mathbf{1}\{g(x,\varepsilon) \leq y\} \mid V = v]$$

and hence the proof is completed by using $\Gamma(x,e) = \mathbf{1}\{g(x,e) \leq y\}$. ■

A.3 Derivation of (3.8)

By Bayes' rule and monotonicity of $v \mapsto h(z,v)$,

$$dF^s_{V,Z}(v,z) = \frac{\mathbb{P}(C > 0 \mid Z = z, V = v) dF_{V,Z}(v,z)}{\mathbb{P}(C > 0)} = \frac{1(h(z,v) > 0) dF_{V,Z}(v,z)}{\int 1(h(z,v) > 0) dF_{V,Z}(v,z)}.$$

The result follows from substitution into (3.7) and using that $F^s_{Y|Z,V}(y \mid z, v) = G(y,x,v)$ by Lemma 1.

B Proofs of Section 4

B.1 Notation

In what follows ϑ denotes a generic value of the control function. It is convenient also to introduce some additional notation, which will be extensively used in the proofs. Let $V_i(\vartheta) := \vartheta(Z_i)$, $W_i(\vartheta) := w(X_i, V_i(\vartheta))$, and $\dot{W}_i(\vartheta) := \partial_v w(X_i, v)|_{v=V_i(\vartheta)}$. When the previous functions are evaluated at the true values we use $V_i = V_i(\vartheta_0)$, $W_i = W_i(\vartheta_0)$, and $\dot{W}_i = \dot{W}_i(\vartheta_0)$. Recall that $A := (Y * 1(C > 0), C, Z, V)$, $T(c) = 1(c \in \overline{\mathcal{C}})$, and $T = T(C)$. For a function $f : \mathcal{A} \mapsto \mathbb{R}$, we use $\|f\|_{T,\infty} = \sup_{a \in \mathcal{A}} |T(c)f(a)|$; for a K-vector of functions $f : \mathcal{A} \mapsto \mathbb{R}^K$, we use $\|f\|_{T,\infty} = \sup_{a \in \mathcal{A}} \|T(c)f(a)\|_2$. We make functions in Υ as well as estimators $\widehat{\vartheta}$ to take values in $[0,1]$. This allows us to simplify notation in what follows.

We adopt the standard notation in the empirical process literature (see, e.g., Van der Vaart, 1998),

$$\mathbb{E}_n[f] = \mathbb{E}_n[f(A)] = n^{-1} \sum_{i=1}^n f(A_i),$$

and

$$\mathbb{G}_n[f] = \mathbb{G}_n[f(A)] = n^{-1/2} \sum_{i=1}^n (f(A_i) - \mathbb{E}_P[f(A)]).$$

When the function $\widehat{f}$ is estimated, the notation should interpreted as:

$$\mathbb{G}_n[\widehat{f}\,] = \mathbb{G}_n[f]\,|_{f=\widehat{f}} \text{ and } \mathbb{E}_P[\widehat{f}\,] = \mathbb{E}_P[f]\,|_{f=\widehat{f}}\,.$$

We also use the concepts of covering entropy and bracketing entropy in the proofs. The

covering entropy $\log N(\epsilon, \mathcal{F}, \|\cdot\|)$ is the logarithm of the minimal number of $\|\cdot\|$-balls of radius ϵ needed to cover the set of functions $\mathcal{F}$. The bracketing entropy $\log N_{[]}(\epsilon, \mathcal{F}, \|\cdot\|)$ is the logarithm of the minimal number of ϵ-brackets in $\|\cdot\|$ needed to cover the set of functions $\mathcal{F}$. An ϵ-bracket $[\ell, u]$ in $\|\cdot\|$ is the set of functions f with $\ell \leq f \leq u$ and $\|u - \ell\| < \epsilon$.

We follow the notation and definitions in Van der Vaart and Wellner (1996) of bootstrap consistency. Let D_n denote the data vector and E_n be the vector of bootstrap weights. Consider the random element $Z_n^b = Z_n(D_n, E_n)$ in a normed space $\mathbb{Z}$. We say that the bootstrap law of Z_n^b consistently estimates the law of some tight random element Z and write $Z_n^b \rightsquigarrow_{\mathbb{P}} Z$ in $\mathbb{Z}$ if

$$\sup_{h \in \mathrm{BL}_1(\mathbb{Z})} \left| \mathbb{E}_P^b h\left(Z_n^b\right) - \mathbb{E}_P h(Z) \right| \to_{\mathbb{P}^b} 0, \tag{B.1}$$

where $\mathrm{BL}_1(\mathbb{Z})$ denotes the space of functions with Lipschitz norm at most 1, $\mathbb{E}_P^b$ denotes the conditional expectation with respect to E_n given the data D_n, and $\to_{\mathbb{P}^b}$ denotes convergence in (outer) probability.

B.2 Proof of Lemma 3

The proof strategy follows closely the argument put forth in Chernozhukov et al. (2015) to deal with the dimensionality and entropy properties of the first step distribution regression estimators.

B.2.1 Auxiliary Lemmas

We start with 2 results on stochastic equicontinuity and a local expansion for the second stage estimators that will be used in the proof of Lemma 3.

Lemma 4 (Stochastic equicontinuity) *Let $\omega \geq 0$ be a positive random variable with $\mathrm{E}_P[\omega] = 1$, $\mathrm{Var}_P[\omega] = 1$, and $\mathbb{E}_P|\omega|^{2+\delta} < \infty$ for some $\delta > 0$, that is independent of $(Y * 1(C > 0), Z, C, V)$, including as a special case $\omega = 1$, and set, for $A = (\omega, Y * 1(C > 0), Z, C, V)$,*

$$f_1(A, \vartheta, \beta) := \omega \cdot [W(\vartheta)^{\mathrm{T}}\beta - Y] \cdot W(\vartheta) \cdot T.$$

Under Assumptions 1–3 the following relations are true.

(a) Consider the set of functions

$$\mathcal{F} = \{f_1(A, \vartheta, \beta)^{\mathrm{T}}\alpha : (\vartheta, \beta) \in \Upsilon_0 \times \mathcal{B}, \alpha \in \mathbb{R}^{\dim(W)}, \|\alpha\|_2 \leq 1\},$$

where $\mathcal{B}$ is a compact set under the $\|\cdot\|_2$ metric containing β_0, Υ_0 is the intersection of Υ, defined in Lemma 2, with a neighborhood of ϑ_0 under the $\|\cdot\|_{T,\infty}$ metric. This class is P-Donsker with a square integrable envelope of the form ω times a constant.

(b) Moreover, if $(\vartheta, \beta) \to (\vartheta_0, \beta_0)$ in the $\|\cdot\|_{T,\infty} \vee \|\cdot\|_2$ metric, then

$$\|f_1(A, \vartheta, \beta) - f_1(A, \vartheta_0, \beta_0)\|_{P,2} \to 0.$$

(c) Hence for any $(\widetilde{\vartheta}, \widetilde{\beta}) \to_{\mathbb{P}} (\vartheta_0, \beta_0)$ in the $\|\cdot\|_{T,\infty} \vee \|\cdot\|_2$ metric such that $\widetilde{\vartheta} \in \Upsilon_0$,

$$\|\mathbb{G}_n f_1(A, \widetilde{\vartheta}, \widetilde{\beta}) - \mathbb{G}_n f_1(A, \vartheta_0, \beta_0)\|_2 \xrightarrow{\mathbb{P}} 0.$$

(d) For for any $(\widehat{\vartheta}, \widetilde{\beta}) \to_{\mathbb{P}} (\vartheta_0, \beta_0)$ *in the* $\|\cdot\|_{T,\infty} \vee \|\cdot\|_2$ *metric so that*

$$\|\widehat{\vartheta} - \widetilde{\vartheta}\|_{T,\infty} = o_{\mathbb{P}}(1/\sqrt{n}), \text{ where } \widetilde{\vartheta} \in \Upsilon_0,$$

we have that

$$\|\mathbb{G}_n f_1(A, \widehat{\vartheta}, \widetilde{\beta}) - \mathbb{G}_n f_1(A, \vartheta_0, \beta_0)\|_2 \to_{\mathbb{P}} 0.$$

Proof of Lemma 4. The proof is divided in subproofs of each of the claims.
Proof of Claim (a). The proof proceeds in several steps.
Step 1. Here we bound the bracketing entropy for

$$\mathcal{I}_1 = \{[W(\vartheta)^{\mathrm{T}}\beta - Y]T : \beta \in \mathcal{B}, \vartheta \in \Upsilon_0\}.$$

For this purpose consider a mesh $\{\vartheta_k\}$ over Υ_0 of $\|\cdot\|_{T,\infty}$ width δ, and a mesh $\{\beta_l\}$ over $\mathcal{B}$ of $\|\cdot\|_2$ width δ. A generic bracket over $\mathcal{I}_1$ takes the form

$$[i_1^0, i_1^1] = [\{W(\vartheta_k)^{\mathrm{T}}\beta_l - \kappa\delta - Y\}T, \{W(\vartheta_k)^{\mathrm{T}}\beta_l + \kappa\delta - Y\}T],$$

where $\kappa = L_W \max_{\beta\in\mathcal{B}} \|\beta\|_2 + L_W$, and $L_W := \|\partial_v w\|_{T,\infty} \vee \|w\|_{T,\infty}$.

Note that this is a valid bracket for all elements of $\mathcal{I}_1$ because for any ϑ located within δ from ϑ_k and any β located within δ from β_l,

$$\begin{aligned} |W(\vartheta)^{\mathrm{T}}\beta - W(\vartheta_k)^{\mathrm{T}}\beta_l|T &\leq |(W(\vartheta) - W(\vartheta_k))^{\mathrm{T}}\beta|T + |W(\vartheta_k)^{\mathrm{T}}(\beta - \beta_l)|T \\ &\leq L_W\delta \max_{\beta\in\mathcal{B}} \|\beta\|_2 + L_W\delta \leq \kappa\delta, \end{aligned} \tag{B.2}$$

and the $\|\cdot\|_{P,2}$-size of this bracket is given by

$$\|i_1^0 - i_1^1\|_{P,2} \leq \sqrt{2}\kappa\delta.$$

Hence, counting the number of brackets induced by the mesh created above, we arrive at the following relationship between the bracketing entropy of $\mathcal{I}_1$ and the covering entropies of Υ_0, and $\mathcal{B}$,

$$\begin{aligned} \log N_{[]}(\epsilon, \mathcal{I}_1, \|\cdot\|_{P,2}) &\lesssim \log N(\epsilon^2, \Upsilon_0, \|\cdot\|_{T,\infty}) + \log N(\epsilon^2, \mathcal{B}, \|\cdot\|_2) \\ &\lesssim 1/(\epsilon^2 \log^4 \epsilon) + \log(1/\epsilon), \end{aligned}$$

and so $\mathcal{I}_1$ is P-Donsker with a constant envelope.
Step 2. Similarly to Step 1, it follows that

$$\mathcal{I}_2 = \{W(\vartheta)^{\mathrm{T}}\alpha T : \vartheta \in \Upsilon_0, \alpha \in \mathbb{R}^{\dim(W)}, \|\alpha\|_2 \leq 1\}$$

also obeys a similar bracketing entropy bound

$$\log N_{[]}(\epsilon, \|\cdot\|_{P,2}) \lesssim 1/(\epsilon^2 \log^4 \epsilon) + \log(1/\epsilon)$$

with a generic bracket taking the form $[i_2^0, i_2^1] = [\{W(\vartheta_k)^{\mathrm{T}}\alpha - \kappa\delta\}T, \{W(\vartheta_k)^{\mathrm{T}}\alpha + \kappa\delta\}T]$. Hence, this class is also P-Donsker with a constant envelope.
Step 3. In this step we verify the claim (a). Note that $\mathcal{F} = \omega \cdot \mathcal{I}_1 \cdot \mathcal{I}_2$. This class has a square-integrable envelope under P. The class $\mathcal{F}$ is P-Donsker by the following argument. Note that the product $\mathcal{I}_1 \cdot \mathcal{I}_2$ of uniformly bounded classes is P-Donsker, e.g., by Theorem 2.10.6 of Van der Vaart and Wellner (1996). Under the stated assumption the final prod-

uct of the random variable ω with the P-Donsker class remains to be P-Donsker by the Multiplier Donsker Theorem, namely Theorem 2.9.2 in Van der Vaart and Wellner (1996).

Proof of Claim (b). The claim follows by the Dominated Convergence Theorem, since any $f_1 \in \mathcal{F}$ is dominated by a square-integrable envelope under P, and, $W(\vartheta)^{\mathrm{T}}\beta T \to W^{\mathrm{T}}\beta_0 T$ and $|W(\vartheta)^{\mathrm{T}}\alpha T - W^{\mathrm{T}}\alpha T| \to 0$ in view of the relation such as (B.2).

Proof of Claim (c). This claim follows from the asymptotic equicontinuity of the empirical process $(\mathbb{G}_n[f_1], f_1 \in \mathcal{F})$ under the $L_2(P)$ metric, and hence also with respect to the $\|\cdot\|_{T,\infty} \vee \|\cdot\|_2$ metric in view of Claim (b).

Proof of Claim (d). It is convenient to set $\widehat{f}_1 := f_1(A, \widehat{\vartheta}, \widetilde{\beta})$ and $\widetilde{f}_1 := f_1(A, \widetilde{\vartheta}, \widetilde{\beta})$. Note that

$$\begin{aligned}
\max_{1\le j\le \dim W} |\mathbb{G}_n[\widehat{f}_1 - \widetilde{f}_1]|_j &\le \max_{1\le j\le \dim W} |\sqrt{n}\mathbb{E}_n[\widehat{f}_1 - \widetilde{f}_1]|_j + \max_{1\le j\le \dim W} |\sqrt{n}\mathbb{E}_P(\widehat{f}_1 - \widetilde{f}_1)|_j \\
&\lesssim \sqrt{n}\mathbb{E}_n[\widehat{\zeta}\,] + \sqrt{n}\mathrm{E}_P[\widehat{\zeta}\,] \\
&\lesssim \mathbb{G}_n[\widehat{\zeta}\,] + 2\sqrt{n}\mathbb{E}_P[\widehat{\zeta}\,],
\end{aligned}$$

where $|f|_j$ denotes the jth element of the application of absolute value to each element of the vector f_1, and $\widehat{\zeta}$ is defined by the following relationship, which holds with probability approaching one,

$$\max_{1\le j\le \dim W} |\widehat{f}_1 - \widetilde{f}_1|_j \lesssim \omega|W(\widehat{\vartheta})^{\mathrm{T}}\widetilde{\beta} - Y|\|W(\widehat{\vartheta}) - W(\widetilde{\vartheta})\|_2 + \omega|W(\widehat{\vartheta})^{\mathrm{T}}\widetilde{\beta} - W(\widetilde{\vartheta})^{\mathrm{T}}\widetilde{\beta}| \lesssim \omega(k + |Y|)\Delta_n =: \widehat{\zeta},$$

where k is a constant such that $k \ge L_W \max_{\beta\in\mathcal{B}} \|\beta\|_2$ with $L_W = \|\partial_v w\|_{T,\infty} \vee \|w\|_{T,\infty}$, and $\Delta_n = o(1/\sqrt{n})$ is a deterministic sequence such that

$$\Delta_n \ge \|\widehat{\vartheta} - \widetilde{\vartheta}\|_{T,\infty}.$$

The second inequality result follows from

$$|W(\widehat{\vartheta})^{\mathrm{T}}\widetilde{\beta} - Y|\|W(\widehat{\vartheta}) - W(\widetilde{\vartheta})\|_2 \lesssim (k + |Y|)\Delta_n, \text{ and } |W(\widehat{\vartheta})^{\mathrm{T}}\widetilde{\beta} - W(\widetilde{\vartheta})^{\mathrm{T}}\widetilde{\beta}| \lesssim k\Delta_n.$$

Then, by part (c) the result follows from

$$\mathbb{G}_n[\widehat{\zeta}\,] = o_{\mathbb{P}}(1), \qquad \sqrt{n}\mathbb{E}_P[\widehat{\zeta}\,] = o_{\mathbb{P}}(1).$$

Indeed,

$$\|\widehat{\zeta}\|_{P,2} \lesssim \sqrt{\mathbb{E}_P\omega^2(k^2 + \mathrm{E}_P(Y^2 \mid C \in \overline{\mathcal{C}}))\Delta_n^2} = o(1) \Rightarrow \mathbb{G}_n[\widehat{\zeta}\,] = o_{\mathbb{P}}(1),$$

and

$$\|\widehat{\zeta}\|_{P,1} \le \mathbb{E}_P|\omega| \cdot (k + \mathbb{E}_P(|Y| \mid C \in \overline{\mathcal{C}})\Delta_n = o(1/\sqrt{n}) \Rightarrow \mathbb{E}_P|\widehat{\zeta}| = o_{\mathbb{P}}(1/\sqrt{n}).$$

Lemma 5 (Local expansion) *Under Assumptions 1–3, for*

$$\begin{aligned}
&\widehat{\delta} = \sqrt{n}(\widetilde{\beta} - \beta_0) = O_{\mathbb{P}}(1); \\
&\widehat{\Delta}(c, r) = \sqrt{n}(\widehat{\vartheta}(c, r) - \vartheta_0(c, r)) = \sqrt{n}\ \mathbb{E}_n[\ell(A, c, r)] + o_{\mathbb{P}}(1) \ \textit{in}\ \ell^\infty(\overline{\mathcal{CR}}), \\
&\|\sqrt{n}\ \mathbb{E}_n[\ell(A, \cdot)]\|_{T,\infty} = O_{\mathbb{P}}(1),
\end{aligned}$$

we have that

$$\sqrt{n}\ \mathbb{E}_P[\{W(\widehat{\vartheta})^{\mathrm{T}}\widetilde{\beta} - Y\}W(\widehat{\vartheta})T] = J\widehat{\delta} + \sqrt{n}\ \mathbb{E}_n\left[f_2(A)\right] + o_{\mathbb{P}}(1),$$

where

$$f_2(a) = \mathbb{E}_P\{[W^{\mathrm{T}}\beta_0 - Y]\dot{W} + W\dot{W}^{\mathrm{T}}\beta_0\}T\ell(a, C, R).$$

Proof of Lemma 5.
Uniformly in $Z \in \overline{\mathcal{Z}}$,

$$\begin{aligned}
&\sqrt{n}\mathbb{E}_P\{W(\widehat{\vartheta})^{\mathrm{T}}\widetilde{\beta} - Y \mid Z\}T \\
&= \sqrt{n}\mathbb{E}_P\{W^{\mathrm{T}}\beta_0 - Y \mid Z\}T + \{W(\bar{\vartheta}_\xi)^{\mathrm{T}}\widehat{\delta} + \dot{W}(\bar{\vartheta}_\xi)^{\mathrm{T}}\bar{\beta}_\xi\widehat{\Delta}(C, R)\}T \\
&= \sqrt{n}\mathbb{E}_P\{W^{\mathrm{T}}\beta_0 - Y \mid Z\}T + \{W^{\mathrm{T}}\widehat{\delta} + \dot{W}^{\mathrm{T}}\beta_0\widehat{\Delta}(C, R)\}T + \rho_Z, \\
&\bar{\rho} = \sup_{\{Z\in\overline{\mathcal{Z}}\}} |\rho_Z| = o_{\mathbb{P}}(1),
\end{aligned}$$

where $\bar{\vartheta}_\xi$ is on the line connecting ϑ_0 and $\widehat{\vartheta}$ and $\bar{\beta}_\xi$ is on the line connecting β_0 and $\widetilde{\beta}$. The first equality follows by the mean value expansion. The second equality follows by uniform continuity of $W(\cdot)$ and $\dot{W}(\cdot)$, $\|\widehat{\vartheta} - \vartheta_0\|_{T,\infty} \xrightarrow{\mathbb{P}} 0$ and $\|\widetilde{\beta} - \beta_0\|_2 \xrightarrow{\mathbb{P}} 0$.

Since the entries of W and $\dot{W}$ are bounded, $\widehat{\delta} = O_{\mathbb{P}}(1)$, and $\|\widehat{\Delta}\|_{T,\infty} = O_{\mathbb{P}}(1)$, with probability approaching one,

$$\begin{aligned}
\sqrt{n}\mathbb{E}_P\{W(\widehat{\vartheta})^{\mathrm{T}}\widetilde{\beta} - Y\}W(\widehat{\vartheta})T &= \mathbb{E}_P\{W^{\mathrm{T}}\beta_0 - Y\}\dot{W}T\widehat{\Delta}(C, R) \\
&\quad + \mathbb{E}_P\{WW^{\mathrm{T}}T\}\widehat{\delta} + \mathbb{E}_P\{W\dot{W}^{\mathrm{T}}\beta_0 T\widehat{\Delta}(C, R)\} + O_{\mathbb{P}}(\bar{\rho}) \\
&= J\widehat{\delta} + \mathbb{E}_P[\{W^{\mathrm{T}}\beta_0 - Y\}\dot{W} + W\dot{W}^{\mathrm{T}}\beta_0]T\widehat{\Delta}(C, R) + o_{\mathbb{P}}(1).
\end{aligned}$$

Substituting in $\widehat{\Delta}(x, r) = \sqrt{n}\ \mathbb{E}_n[\ell(A, x, r)] + o_{\mathbb{P}}(1)$ and interchanging $\mathbb{E}_P$ and $\mathbb{E}_n$,

$$\mathbb{E}_P[\{W^{\mathrm{T}}\beta_0 - Y\}\dot{W} + W\dot{W}^{\mathrm{T}}\beta_0]T\widehat{\Delta}(C, R) = \sqrt{n}\ \mathbb{E}_n[g(A)] + o_{\mathbb{P}}(1),$$

since $\|[\{W^{\mathrm{T}}\beta_0 - Y\}\dot{W} + \ W\dot{W}^{\mathrm{T}}\beta_0]T\|_{P,2}$ is bounded . The claim of the lemma follows.

B.2.2 Proof of Lemma 3.

The proof is divided in two parts corresponding to the CLT and bootstrap CLT.

CLT: In this part we show $\sqrt{n}(\widehat{\beta} - \beta_0) \rightsquigarrow J^{-1}G$ in $\mathbb{R}^{d_w}$.
Step 1. This step shows that $\sqrt{n}(\widehat{\beta} - \beta_0) = O_{\mathbb{P}}(1)$. Recall that

$$\widehat{\beta} = \arg\min_{\beta\in\mathbb{R}^{d_w}} \mathbb{E}_n[(Y - W(\widehat{\vartheta})^{\mathrm{T}}\beta)^2 T].$$

Due to convexity of the objective function, it suffices to show that for any $\epsilon > 0$ there exists a finite positive constant B_ϵ such that ,

$$\liminf_{n\to\infty} \mathbb{P}\left(\inf_{\|\eta\|_2=1} \sqrt{n}\eta^{\mathrm{T}}\mathbb{E}_n\left[\widehat{f}_{1,\eta,B_\epsilon}\right] > 0\right) \geq 1 - \epsilon, \tag{B.3}$$

where

$$\widehat{f}_{1,\eta,B_\epsilon}(A) := \left\{W(\widehat{\vartheta})^{\mathrm{T}}(\beta_0 + B_\epsilon\eta/\sqrt{n}) - Y\right\}W(\widehat{\vartheta})T.$$

Let

$$f_1(A) := \left\{W^{\mathrm{T}}\beta_0 - Y\right\} WT.$$

Then uniformly in $\|\eta\|_2 = 1$,

$$\begin{aligned}
\sqrt{n}\eta^{\mathrm{T}}\mathbb{E}_n[\widehat{f}_{1,\eta,B_\epsilon}] &= \eta^{\mathrm{T}}\mathbb{G}_n[\widehat{f}_{1,\eta,B_\epsilon}] + \sqrt{n}\eta^{\mathrm{T}}\mathbb{E}_P[\widehat{f}_{1,\eta,B_\epsilon}] \\
&=_{(1)} \eta^{\mathrm{T}}\mathbb{G}_n[f_1] + o_{\mathbb{P}}(1) + \eta^{\mathrm{T}}\sqrt{n}\mathbb{E}_P[\widehat{f}_{1,\eta,B_\epsilon}] \\
&=_{(2)} \eta^{\mathrm{T}}\mathbb{G}_n[f_1] + o_{\mathbb{P}}(1) + \eta^{\mathrm{T}}J\eta B_\epsilon + \eta^{\mathrm{T}}\mathbb{G}_n[f_2] + o_{\mathbb{P}}(1) \\
&=_{(3)} O_{\mathbb{P}}(1) + o_{\mathbb{P}}(1) + \eta^{\mathrm{T}}J\eta B_\epsilon + O_{\mathbb{P}}(1) + o_{\mathbb{P}}(1),
\end{aligned}$$

where relations (1) and (2) follow by Lemma 4 and Lemma 5 with $\widetilde{\beta} = \beta_0 + B_\epsilon\eta/\sqrt{n}$, respectively, using that $\|\widehat{\vartheta} - \widetilde{\vartheta}\|_{T,\infty} = o_{\mathbb{P}}(1/\sqrt{n})$, $\widetilde{\vartheta} \in \Upsilon$, $\|\widetilde{\vartheta} - \vartheta_0\|_{T,\infty} = O_{\mathbb{P}}(1/\sqrt{n})$ and $\|\beta_0 + B_\epsilon\eta/\sqrt{n} - \beta_0\|_2 = O(1/\sqrt{n})$; relation (3) holds because f_1 and f_2 are P-Donsker by step-2 below. Since J is positive definite, with minimal eigenvalue bounded away from zero, the inequality (B.3) follows by choosing B_ϵ as a sufficiently large constant.
Step 2. In this step we show the main result. Let

$$\widehat{f}_1(A) := \left\{W(\widehat{\vartheta})^{\mathrm{T}}\widehat{\beta} - Y\right\} W(\widehat{\vartheta})T.$$

From the first order conditions of the least squares problem,

$$\begin{aligned}
0 = \sqrt{n}\mathbb{E}_n\left[\widehat{f}_1\right] &= \mathbb{G}_n\left[\widehat{f}_1\right] + \sqrt{n}\mathbb{E}_P\left[\widehat{f}_1\right] \\
&=_{(1)} \mathbb{G}_n[f_1] + o_{\mathbb{P}}(1) + \sqrt{n}\mathbb{E}_P\left[\widehat{f}_1\right] \\
&=_{(2)} \mathbb{G}_n[f_1] + o_{\mathbb{P}}(1) + J\sqrt{n}(\widehat{\beta} - \beta_0) + \mathbb{G}_n[f_2] + o_{\mathbb{P}}(1),
\end{aligned}$$

where relations (1) and (2) follow by Lemma 4 and Lemma 5 with $\widetilde{\beta} = \widehat{\beta}$, respectively, using that $\|\widehat{\vartheta} - \widetilde{\vartheta}\|_{T,\infty} = o_{\mathbb{P}}(1/\sqrt{n})$, $\widetilde{\vartheta} \in \Upsilon$, and $\|\widetilde{\vartheta} - \vartheta\|_{T,\infty} = O_{\mathbb{P}}(1/\sqrt{n})$ by Lemma 2, and $\|\widehat{\beta} - \beta_0\|_2 = O_{\mathbb{P}}(1/\sqrt{n})$.

Therefore by invertibility of J,

$$\sqrt{n}(\widehat{\beta} - \beta_0) = -J^{-1}\mathbb{G}_n(f_1 + f_2) + o_{\mathbb{P}}(1).$$

By the Central Limit Theorem

$$\mathbb{G}_n(f_1 + f_2) \rightsquigarrow G \text{ in } \mathbb{R}^{d_w}, \quad G \sim N(0, \Omega), \quad \Omega = \mathbb{E}_P[(f_1 + f_2)(f_1 + f_2)^{\mathrm{T}}],$$

where Ω is specified in the lemma. Conclude that

$$\sqrt{n}(\widehat{\beta} - \beta_0) \rightsquigarrow J^{-1}G \text{ in } \mathbb{R}^{d_w}.$$

Bootstrap CLT: In this part we show $\sqrt{n}(\widehat{\beta}^b - \widehat{\beta}) \rightsquigarrow_{\mathbb{P}} J^{-1}G$ in $\mathbb{R}^{d_w}$.
Step 1. This step shows that $\sqrt{n}(\widehat{\beta}^b - \beta_0) = O_{\mathbb{P}}(1)$ under the unconditional probability $\mathbb{P}$. Recall that

$$\widehat{\beta}^b = \arg\min_{\beta \in \mathbb{R}^{\dim(W)}} \mathbb{E}_n[\omega(Y - W(\widehat{\vartheta}^b)^{\mathrm{T}}\beta)^2 T],$$

where ω is the random variable used in the weighted bootstrap. Due to convexity of the objective function, it suffices to show that for any $\epsilon > 0$ there exists a finite positive

constant B_ϵ such that

$$\liminf_{n\to\infty} \mathbb{P}\left(\inf_{\|\eta\|_2=1} \sqrt{n}\eta^{\mathrm{T}}\mathbb{E}_n\left[\widehat{f}^b_{1,\eta,B_\epsilon}\right] > 0\right) \geq 1-\epsilon, \tag{B.4}$$

where

$$\widehat{f}^b_{1,\eta,B_\epsilon}(A) := \omega \cdot \left\{[W(\widehat{\vartheta}^b)^{\mathrm{T}}(\beta_0 + B_\epsilon\eta/\sqrt{n})] - Y\right\} W(\widehat{\vartheta}^b)T.$$

The result then follows by an analogous argument to step 1 in the proof of the CLT, which we do not repeat here.

Step 2. In this step we show that $\sqrt{n}(\widehat{\beta}^b - \beta_0) = -J^{-1}\mathbb{G}_n(f^b_1 + f^b_2) + o_{\mathbb{P}}(1)$ under the unconditional probability $\mathbb{P}$. Let

$$\widehat{f}^b_1(A) := \omega \cdot \{W(\widehat{\vartheta}^b)^{\mathrm{T}}\widehat{\beta}^b - Y\}W(\widehat{\vartheta}^b)T.$$

From the first order conditions of the least squares problem in the weighted sample,

$$\begin{aligned} 0 = \sqrt{n}\mathbb{E}_n\left[\widehat{f}^b_1\right] &= \mathbb{G}_n\left[\widehat{f}^b_1\right] + \sqrt{n}\mathbb{E}_P\left[\widehat{f}^b_1\right] \\ &=_{(1)} \mathbb{G}_n[f^b_1] + o_{\mathbb{P}}(1) + \sqrt{n}\mathbb{E}_P\left[\widehat{f}^b_1\right] \\ &=_{(2)} \mathbb{G}_n[f^b_1] + o_{\mathbb{P}}(1) + J\sqrt{n}(\widehat{\beta}^b - \beta_0) + \mathbb{G}_n[f^b_2] + o_{\mathbb{P}}(1), \end{aligned}$$

where relations (1) and (2) follow by Lemma 4 and Lemma 5 with $\widetilde{\beta} = \widehat{\beta}^b$, respectively, using that $\|\widehat{\vartheta}^b - \widetilde{\vartheta}^b\|_{T,\infty} = o_{\mathbb{P}}(1/\sqrt{n})$, $\widetilde{\vartheta}^b \in \Upsilon$ and $\|\widetilde{\vartheta}^b - \vartheta_0\|_{T,\infty} = O_{\mathbb{P}}(1/\sqrt{n})$ by Lemma 2, and $\|\widehat{\beta}^b - \beta_0\|_2 = O_{\mathbb{P}}(1/\sqrt{n})$. Therefore by invertibility of J,

$$\sqrt{n}(\widehat{\beta}^b - \beta_0) = -J^{-1}\mathbb{G}_n(f^b_1 + f^b_2) + o_{\mathbb{P}}(1).$$

Step 3. In this final step we establish the behavior of $\sqrt{n}(\widehat{\beta}^b - \widehat{\beta})$ under $\mathbb{P}^b$. Note that $\mathbb{P}^b$ denotes the conditional probability measure, namely the probability measure induced by draws of $\omega_1, \ldots, \omega_n$ conditional on the data $A_1, ..., A_n$. By Step 2 of the proof of the CLT and Step 2 of the proof of the bootstrap CLT, we have that under $\mathbb{P}$:

$$\sqrt{n}(\widehat{\beta}^b - \beta_0) = -J^{-1}\mathbb{G}_n(f^b_1 + f^b_2) + o_{\mathbb{P}}(1),\ \sqrt{n}(\widehat{\beta} - \beta_0) = -J^{-1}\mathbb{G}_n(f_1 + f_2) + o_{\mathbb{P}}(1).$$

Hence, under $\mathbb{P}$

$$\sqrt{n}(\widehat{\beta}^b - \widehat{\beta}) = -J^{-1}\mathbb{G}_n(f^b_1 - f_1 + f^b_2 - f_2) + r_n = -J^{-1}\mathbb{G}_n((\omega - 1)(f_1 + f_2)) + r_n,$$

where $r_n = o_{\mathbb{P}}(1)$. Note that it is also true that

$$r_n = o_{\mathbb{P}^b}(1) \text{ in } \mathbb{P}\text{-probability},$$

where the latter statement means that for every $\epsilon > 0$, $\mathbb{P}^b(\|r_n\|_2 > \epsilon) = o_{\mathbb{P}}(1)$. Indeed, this follows from Markov inequality and by

$$\mathbb{E}_{\mathbb{P}}[\mathbb{P}^b(\|r_n\|_2 > \epsilon)] = \mathbb{P}(\|r_n\|_2 > \epsilon) = o(1),$$

where the latter holds by the Law of Iterated Expectations and $r_n = o_{\mathbb{P}}(1)$.

Note that $f^b_1 = \omega \cdot f_1$ and $f^b_2 = \omega \cdot f_2$, where f_1 and f_2 are P-Donsker by step-2 of the proof of the first part and $\mathbb{E}_P\omega^2 < \infty$. Then, by the Conditional Multiplier Central Limit

Theorem, e.g., Lemma 2.9.5 in Van der Vaart and Wellner (1996),

$$G_n^b := \mathbb{G}_n((\omega - 1)(f_1 + f_2)) \rightsquigarrow_{\mathbb{P}} G \text{ in } \mathbb{R}^{d_w}.$$

Conclude that

$$\sqrt{n}(\widehat{\beta}^b - \widehat{\beta}) \rightsquigarrow_{\mathbb{P}} J^{-1}G \text{ in } \mathbb{R}^{d_w}.$$

B.3 Proof of Theorem 5

In this section we use the notation $W_x(\vartheta) = w(x, V(\vartheta))$ such that $W_x = w(x, V(\vartheta_0))$.

We focus on the proof for the estimator of the ASF, because the proof for the estimator of the ASF on the treated can be obtained by analogous arguments. The results for the estimator of the ASF follow by a similar argument to the proof of Lemma 3 using Lemmas 6 and 7 in place of Lemmas 4 and 5, and the delta method. For the sake of brevity, here we just outline the proof of the FCLT.

Let $\psi_x(A, \vartheta, \beta) := W_x(\vartheta)^{\mathrm{T}}\beta T$ such that $\mu_S(x) = \mathbb{E}_P\psi_x(A, \vartheta_0, \beta_0)/\mathbb{E}_P T$ and $\widehat{\mu}_S(x) = \mathbb{E}_n\psi_x(A, \widehat{\vartheta}, \widehat{\beta})/\mathbb{E}_n T$. Then, for $\widehat{\psi}_x := \psi_x(A, \widehat{\vartheta}, \widehat{\beta})$ and $\psi_x := \psi_x(A, \vartheta_0, \beta_0)$,

$$\begin{aligned}
\sqrt{n}\left[\mathbb{E}_n\psi_x(A, \widehat{\vartheta}, \widehat{\beta}) - \mathbb{E}_P\psi_x(A, \vartheta_0, \beta_0)\right] &= \mathbb{G}_n\left[\widehat{\psi}_x\right] + \sqrt{n}\mathbb{E}_P\left[\widehat{\psi}_x - \psi_x\right] \\
&=_{(1)} \mathbb{G}_n[\psi_x] + o_{\mathbb{P}}(1) + \sqrt{n}\mathbb{E}_P\left[\widehat{\psi}_x - \psi_x\right] \\
&=_{(2)} \mathbb{G}_n[\psi_x] + o_{\mathbb{P}}(1) + \mathbb{G}_n[\sigma_x] + o_{\mathbb{P}}(1),
\end{aligned}$$

where relations (1) and (2) follow by Lemma 6 and Lemma 7 with $\widetilde{\beta} = \widehat{\beta}$, respectively, using that $\|\widehat{\vartheta} - \widetilde{\vartheta}\|_{T,\infty} = o_{\mathbb{P}}(1/\sqrt{n})$, $\widetilde{\vartheta} \in \Upsilon$, and $\|\widetilde{\vartheta} - \vartheta\|_{T,\infty} = O_{\mathbb{P}}(1/\sqrt{n})$ by Lemma 2, and $\sqrt{n}(\widehat{\beta} - \beta_0) = -J^{-1}\mathbb{G}_n(f_1 + f_2) + o_{\mathbb{P}}(1)$ from step 2 of the proof of Lemma 3.

The functions $x \mapsto \psi_x$ and $x \mapsto \sigma_x$ are P-Donsker by Example 19.7 in Van der Vaart (1998) because they are Lipschitz continuous on $\overline{\mathcal{X}}$. Hence, by the Functional Central Limit Theorem

$$\mathbb{G}_n(\psi_x + \sigma_x) \rightsquigarrow Z(x) \text{ in } \ell^\infty(\overline{\mathcal{X}}),$$

where $x \mapsto Z(x)$ is a zero mean Gaussian process with uniformly continuous sample paths and covariance function

$$\mathrm{Cov}_P[\psi_{x_0} + \sigma_{x_0}, \psi_{x_1} + \sigma_{x_1}], \quad x_0, x_1 \in \overline{\mathcal{X}}.$$

The result follows by the functional delta method applied to the ratio of $\mathbb{E}_n\psi_x(A, \widehat{\vartheta}, \widehat{\beta})$ and $\mathbb{E}_n T$ using that

$$\begin{pmatrix} \mathbb{G}_n\psi_x(A, \widehat{\vartheta}, \widehat{\beta}) \\ \mathbb{G}_n T \end{pmatrix} \rightsquigarrow \begin{pmatrix} Z(x) \\ Z_T \end{pmatrix},$$

where $Z_T \sim N(0, p_T(1 - p_T))$,

$$\mathrm{Cov}_P(Z(x), Z_T) = G_T(x)p_T(1 - p_T),$$

and

$$\begin{aligned}
&\mathrm{Cov}_P[\psi_{x_0} + h_{x_0}, \psi_{x_1} + \sigma_{x_1} \mid T = 1] \\
&\qquad = \frac{\mathrm{Cov}_P[\psi_{x_0} + \sigma_{x_0}, \psi_{x_1} + \sigma_{x_1}] - \mu_T(x_0)\mu_T(x_1)p_T(1 - p_T)}{p_T}.
\end{aligned}$$

Lemma 6 (Stochastic equicontinuity) *Let $\omega \geq 0$ be a positive random variable with $\mathbb{E}_P[\omega] = 1$, $\mathrm{Var}_P[\omega] = 1$, and $\mathbb{E}_P|\omega|^{2+\delta} < \infty$ for some $\delta > 0$, that is independent of $(Y1(C > 0), C, Z, V)$, including as a special case $\omega = 1$, and set, for $A = (\omega, Y1(C > 0), C, Z, V)$,*

$$\psi_x(A, \vartheta, \beta) := \omega \cdot W_x(\vartheta)^{\mathrm{T}} \beta \cdot T.$$

Under Assumptions 1–3, the following relations are true.

(a) Consider the set of functions

$$\mathcal{F} := \{\psi_x(A, \vartheta, \beta) : (\vartheta, \beta, x) \in \Upsilon_0 \times \mathcal{B} \times \overline{\mathcal{X}}\},$$

where $\overline{\mathcal{X}}$ is a compact subset of $\mathbb{R}$, $\mathcal{B}$ is a compact set under the $\|\cdot\|_2$ metric containing β_0, Υ_0 is the intersection of Υ, defined in Lemma 2, with a neighborhood of ϑ_0 under the $\|\cdot\|_{T,\infty}$ metric. This class is P-Donsker with a square integrable envelope of the form ω times a constant.

(b) Moreover, if $(\vartheta, \beta) \to (\vartheta_0, \beta_0)$ in the $\|\cdot\|_{T,\infty} \vee \|\cdot\|_2$ metric , then

$$\sup_{x \in \overline{\mathcal{X}}} \|\psi_x(A, \vartheta, \beta) - \psi_x(A, \vartheta_0, \beta_0)\|_{P,2} \to 0.$$

(c) Hence for any $(\widetilde{\vartheta}, \widetilde{\beta}) \to_{\mathbb{P}} (\vartheta_0, \beta_0)$ in the $\|\cdot\|_{T,\infty} \vee \|\cdot\|_2$ metric such that $\widetilde{\vartheta} \in \Upsilon_0$,

$$\sup_{x \in \overline{\mathcal{X}}} \|\mathbb{G}_n \psi_x(A, \widetilde{\vartheta}, \widetilde{\beta}) - \mathbb{G}_n \psi_x(A, \vartheta_0, \beta_0)\|_2 \to_{\mathbb{P}} 0.$$

(d) For for any $(\widehat{\vartheta}, \widetilde{\beta}) \to_{\mathbb{P}} (\vartheta_0, \beta_0)$ in the $\|\cdot\|_{T,\infty} \vee \|\cdot\|_2$ metric , so that

$$\|\widehat{\vartheta} - \widetilde{\vartheta}\|_{T,\infty} = o_{\mathbb{P}}(1/\sqrt{n}), \text{ where } \widetilde{\vartheta} \in \Upsilon_0,$$

we have that

$$\sup_{x \in \overline{\mathcal{X}}} \|\mathbb{G}_n \psi_x(A, \widehat{\vartheta}, \widetilde{\beta}) - \mathbb{G}_n \psi_x(A, \vartheta_0, \beta_0)\|_2 \to_{\mathbb{P}} 0.$$

Proof of Lemma 6. The proof is omitted because it is similar to the proof of Lemma 4.

Lemma 7 (Local expansion) *Under Assumptions 1–3, for*

$$\begin{aligned}
&\widehat{\delta} = \sqrt{n}(\widetilde{\beta} - \beta_0) = O_{\mathbb{P}}(1); \\
&\widehat{\Delta}(x, r) = \sqrt{n}(\widehat{\vartheta}(x, r) - \vartheta_0(x, r)) = \sqrt{n}\ \mathbb{E}_n[\ell(A, x, r)] + o_{\mathbb{P}}(1) \text{ in } \ell^\infty(\overline{\mathcal{X}\mathcal{R}}), \\
&\|\sqrt{n}\ \mathbb{E}_n[\ell(A, \cdot)]\|_{T,\infty} = O_{\mathbb{P}}(1),
\end{aligned}$$

we have that

$$\sqrt{n}\ \left\{\mathbb{E}_P W_x(\widehat{\vartheta})^{\mathrm{T}} \widetilde{\beta} T - \mathbb{E}_P W_x^{\mathrm{T}} \beta_0 T\right\} = \mathbb{E}_P\{W_x T\}^{\mathrm{T}} \widehat{\delta} + \mathbb{E}_P\{\dot{W}_x^{\mathrm{T}} \beta_0 T \ell(a, X, R)\}\big|_{a=A} + \bar{o}_{\mathbb{P}}(1),$$

where $\bar{o}_{\mathbb{P}}(1)$ denotes order in probability uniform in $x \in \overline{\mathcal{X}}$.

Proof of Lemma 7. The proof is omitted because it is similar to the proof of Lemma 5.